1-1-2006

Equating high stakes educational measurements : a study of design and consequences.

Bob Wajizigha Chulu
University of Massachusetts Amherst

EQUATING HIGH STAKES EDUCATIONAL MEASUREMENTS: A STUDY OF DESIGN AND CONSEQUENCES

A Dissertation Presented

by

Bob Wajizigha Chulu

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF EDUCATION

May 2006

School of Education

UMI Number: 3215910

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

EQUATING HIGH STAKES EDUCATIONAL MEASUREMENTS: A STUDY DESIGN AND CONSEQUENCES

A Dissertation Presented

by

Bob Wajizigha Chulu

Approved as to style and content by:

_____
Stephen G. Sireci, Chair

_____
Ronald K. Hambleton, Member

_____
Craig S. Wells, Member

_____
Aline G. Sayer, Member

_____
Christine McCormick, Dean
School of Education

EQUATING HIGH STAKES EDUCATIONAL MEASUREMENTS: A STUDY OF DESIGN AND CONSEQUENCES

A Dissertation Presented

by

Bob Wajizigha Chulu

Approved as to style and content by:

Stephen G. Sireci, Chair

Ronald K. Hambleton, Member

Craig S. Wells, Member

Aline G. Sayer, Member

Christine McCormick, Dean
School of Education

DEDICATION

To my parents, wife, and kids.

ACKNOWLEDGEMENT

Many people contributed their talents to this work. First, I wish to thank my committee for their invaluable support and guidance that made the completion of this dissertation possible. I am especially grateful to my advisor, Professor Stephen Sireci who guided me at every step along the way. His insightful improvements on my initially vague research ideas, willingness to read my drafts with grace, and more importantly, the many hours he spent with me in dialogue as we refined my thinking on equating, have vastly enriched this dissertation.

The impetus for this work came from Professor Ronald Hambleton who gave us the chance to work with real data for class assignments. It was through these class assignments that I developed interest in equating and acquired skills that made this dissertation easy to handle. I cannot fail to thank Professor Craig Wells, who was always available and tolerated my numerous questions without prior appointments. Furthermore, equating is a statistical process and I would not have successfully completed the dissertation without Professor Lisa Keller, who inculcated these important statistical skills into my schema. I am equally grateful to Professor Aline Sayer for accepting to serve on my committee. As always, it was a joy to work with these great professors.

My appreciation also goes to the Center for International Education (CIE) for funding part of this work under the UPIC Project and to Professor Ronald Hambleton for providing the financial support to complement what UPIC offered. I am grateful to students and staff in Zomba Schools who participated in this study. Students and Staff in CIE and the Research and Evaluation Methods Program (REMP) also provided the community I could always go to when I need anything. I am very grateful to them.

More importantly, my deep appreciation and gratitude go to my loving family. My wife, Chrissy, who single-handedly labored to take care of our children, and without her love and understanding it would be hard to continue with the program. Our two wonderful kids, Wongani and Upendo who missed their dad, but always remembered to sing school songs to him over the phone provided the cheerful atmosphere I needed to keep me going. My dear friend Ellen Krause and the whole Cornerstone Chapel supported me spiritually and morally, opening their homes to me when I felt lonely, and attending to me when I am sick. God bless you more abundantly.

Finally, I would like to acknowledge the incalculable debt I owe to my parents, Mackson and Margret Chulu who asked me to get educated on their behalf. People ridiculed them for sacrificing everything they had to support my education, but they never retreated. Their courage, zest for life, love of learning, and strength of spirit have taught me far more about how to develop my skills than any academic study could ever convey. I hope that I have fulfilled their dream.

ABSTRACT

EQUATING HIGH STAKES EDUCATIONAL MEASUREMENTS: A STUDY OF DESIGN AND CONSEQUENCES

MAY 2006

BOB WAJIZIGHA CHULU, B.Ed., UNIVERSITY OF MALAWI, CHANCELLOR COLLEGE

M.Ed., UNIVERSITY OF MASSACHUSETTS AMHERST

Ed.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by Professor Stephen G. Sireci

The practice of equating educational and psychological tests to create comparable and interchangeable scores is increasingly becoming appealing to most testing and credentialing agencies. However, the Malawi National Examinations Board (MANEB) and many other testing organizations in Africa and Europe do not conduct equating and the consequences of not equating tests have not been clearly documented. Furthermore, there are no proper equating designs for some agencies to employ because they administer tests annually to different examinee' populations and they disclose all items after each administration. Therefore, the purposes of this study were to: (1) determine whether it was necessary to equate MANEB tests; (2) investigate consequences of not equating educational tests; and (3) explore the possibility of using an external anchor test that is administered separately from the target tests to equate scores.

The study used 2003, 2004, and 2005 Primary School Leaving Certificate (PSLCE) Mathematics scores for two randomly equivalent groups of eighth grade examinees drawn from 12 primary schools in the Zomba district in Malawi. In the first administration, group A took the 2004 test while group B took the 2003 form. In the

second administration both groups took an external anchor test and five weeks later, they both took the 2005 test. Data were analyzed using identity and log-linear methods, t-tests, decision consistency analyses, classification consistency analyses, and by computing reduction in uncertainty, and the root mean square difference indices. Both linear and post-smoothed equipercentile methods were used to equate test scores.

The study revealed that: (1) score distributions and test difficulties were dissimilar across test forms signifying that equating is necessary; (2) classification of students into grade categories across forms were different before equating, but similar after equating; and (3) the external anchor test design performed in the same way as the random groups design.

The results suggest that MANEB should equate tests scores to improve consistency of decisions and to match their distributions and difficulty levels across forms. Given the current policy of exam discloser, the use of an external anchor test that is administered separately from the operational form to equate score is recommended.

TABLE OF CONTENTS

Page

ACKNOWLEDGEMENTS..... v

ABSTRACT..... vii

LIST OF TABLES..... xiii

LIST OF FIGURES..... xv

CHAPTER

1. INTRODUCTION..... 1

1.1 Definition and Conditions for Equating.....2
1.2 Rationale for Equating..... 3
1.3 Equating MANEB Tests..... 4

1.3.1 Is it Necessary to Equate MANEB Tests.....4
1.3.2 Scaling Educational Tests..... 5
1.3.3 Standards of Examinations..... 7
1.3.4 Lack of Appropriate Equating Design..... 7

1.4 Statement of the Problem.....8
1.5 Purpose of the Study..... 9

2. LITERATURE REVIEW..... 11

2.1 Determining Whether to Equate..... 11

2.1.1 Is Equating Appropriate?..... 12
2.1.2 Is Equating Necessary?..... 14

2.2 Steps for Implementing Equating..... 15
2.3 Equating Designs..... 17

2.3.1 Single Group (SG) Design..... 17
2.3.2 Counterbalanced (CB) Design..... 18
2.3.3 Random Groups (RG) Design..... 18
2.3.4 Non-Equivalent Groups with Anchor Test (NEAT) Design..... 19

2.4 Equating Methods..... 20

2.4.1 Mean Equating.....22
2.4.2 Linear Equating.....23
2.4.3 Equipercentile Equating.....24
2.4.4 Other Equating Designs.....26

2.5 Equating Using Anchor Test.....27

2.5.1 Using Common Items as Anchor.....28
2.5.2 Using a Different test as Anchor.....31
2.5.3 Using Other variables as Anchor.....32

2.6 Criteria for Evaluating the Adequacy of Equating.....34

2.6.1 Standard Error of Equating.....35
2.6.2 Consequences of Equating.....37
2.6.3 The Root Mean Square Difference.....38
2.6.4 The Reduction in Uncertainty Index.....39

2.7 Summary of the Review.....40

3. METHOD.....44

3.1 Participants.....44

3.1.1 Schools and Students.....44
3.1.2 Teachers.....45
3.1.3 MANEB Officials.....46

3.2 Instruments.....46

3.2.1 Test Forms.....46
3.2.2 External Anchor Test.....47
3.2.3 Survey Items.....48

3.3 Data Collection.....49

3.3.1 Data Collection Design.....49
3.3.2 Administration of the Tests.....50
3.3.3 Scoring.....51
3.3.4 Setting Cut Scores.....52

3.4 Preliminary Analyses.....52

3.4.1 Item and Reliability Analyses.....53
3.4.2 Choosing a Smoothing Procedure.....53
3.4.3 Students' Motivation.....55

3.4.4 Establishing Group Equivalency.....56

3.5 Data Analysis.....56

3.5.1 Is it Necessary to Equate these Tests?.....57

3.5.1.1 Comparing the Difficulty of the Tests.....57
3.5.1.2 Comparing Score Distributions of the Tests.....57

3.5.2 Invariance of Examination Standards.....58
3.5.3 Effect of Equating on Examinees Classification.....59
3.5.4 Equating Using External Anchor Test.....59

3.5.4.1 Were there Significant Learning Effects.....60
3.5.4.2 How Useful is the Anchor Test.....60
3.5.4.3 Equating via the Anchor Test.....61

3.6 Summary of the Method.....62

4. RESULTS.....64

4.1 Results from Preliminary Analyses.....65

4.1.1 Item and Reliability Analyses.....65

4.1.1.1 Item Discrimination.....65
4.1.1.2 Item Difficult.....66
4.1.1.3 Reliability.....67

4.1.2 Choice of the Smoothing Models.....68

4.1.2.1 Choosing a Log-linear Model.....68
4.1.2.2 Fitting the Beta4 Compound Binomial Model.....69
4.1.2.3 Choosing the Cubic Spline Model.....69
4.1.2.4 Comparing Smoothing Models.....70

4.1.3 Level of Motivation.....71
4.1.4 Establishing Group Equivalency.....72

4.1.4.1 Comparing Performance on External Anchor Test.....72
4.1.4.2 Comparing Performance on 2005 Test.....73

4.2 Is it Necessary to Equate these Tests.....73

4.2.1 Comparing the Difficulty of Tests.....73
4.2.2 Comparing Score Distributions.....74

4.3 Invariance of Examination Standards..... 75
4.4 Effects on Examinees' Classification..... 77

4.4.1 Pass rates on Test Forms..... 77
4.4.2 Classification of Candidates into Grade..... 78
4.4.3 Decision Consistency..... 78

4.5 Equating Using External Anchor Test..... 79

4.5.1 Ruling Out Learning Effects..... 80
4.5.2 How Useful is the Anchor Test?..... 81

4.5.2.1 Reduction in Uncertainty Index..... 81
4.5.2.2 Comparing Equating Designs..... 82

4.5.3 Equating via the Anchor Test..... 83

4.5.3.1 Conversion Tables..... 83
4.5.3.2 Mean Square Equating Errors..... 83
4.5.3.3 Comparing Students' Classification..... 84

4.6 Summary of results..... 85

5. SUMMARY OF FINDINGS..... 114

5.1 Summary of Findings..... 114

5.1.1 Is it Necessary to Equate MANEB Tests..... 114
5.1.2 Consequences of not Equating Educational Tests..... 116
5.1.3 Equating Using External Anchor Test..... 121

5.2 Significance of the Findings..... 125
5.3 Delimitations and Direction for Future Research..... 126
5.4 Recommendations..... 128

REFERENCES..... 130

LIST OF TABLES

Table		Page
3.1	Survey Questions	48
3.2	Data Collection Design	49
4.1	Item-Total Correlation, Mean and Alpha Values for 2004 and 2003 Tests	87
4.2	Items-Total Correlation, Mean and Alpha Values for Anchor Test	88
4.3	Correlation Coefficients Between Tests and Subtests	88
4.4	Moments for Presmoothing Score Distributions	89
4.5	Moments for Postsmoothed Score Distributions	90
4.6	Descriptive Statistics for Survey Questions	90
4.7	Regression Statistics for Survey Questions	91
4.8	Descriptive Statistics of Scores on Anchor Test	91
4.9	Descriptive Statistics of Scores on Test Forms	92
4.10	Group Differences on Test Forms	92
4.11	Unsmoothed Raw-to-Raw Score Conversion Table for 2004 to 2003 Scores	93
4.12	Smoothed Raw-to-Raw Score Conversion Table for 2004 to 2003 Scores	94
4.13	Smoothed Raw-to-Raw Score Conversion Table for 2003 to 2004 Scores	95
4.14	Grade Boundaries for 2004 Test Form	96
4.15	Grade Boundaries for 2003 Test Form	96
4.16	Operated and Equated Cut Scores	96
4.17	Standard (Z-Scores) Scores	97

4.18 Pass Rates on 2004 and 2003 Test Forms.....98

4.19 Classification of Candidates Using Cut Scores on Reference (2003) Form.....98

4.20 Decision Consistency for 2005 and 2004 Tests Before Equating.....99

4.21 Decision Consistency for 2005 and 2004 Test After Equating.....99

4.22 Decision Consistency for 2005 and 2003 Tests Before Equating.....100

4.23 Decision Consistency for 2005 and 2003 Tests After Equating.....100

4.24 Reduction in Uncertainty Indices (RIU).....101

4.25 Standardized Root Mean Square Differences (RMSDs) and Mean Square Equating Error (MSEE).....101

4.26 Equipercentile Raw-to-Raw Score Conversion Tables for 2005 to 2004 Scores.....102

4.27 Equipercentile Raw-to-Raw Score Conversion Tables for 2005 to 2003 Scores.....103

4.28 Linear Raw-to-Raw Score Conversion Tables for 2005 to 2004 Scores.....104

4.29 Linear Raw-to-Raw Score Conversion Tables for 2005 to 2003 Scores.....105

4.30 Pass Rates on 2005, 2004 and 2003 Tests before and after Equipercentile Equating.....106

4.31 Pass Rates on 2005, 2004 and 2003 Test before and after Tucker Linear Equating.....106

LIST OF FIGURES

Figure		Page
4.1	Unsmoothed Function versus C = 1 Smoothed Function	107
4.2	Unsmoothed Function versus C = 4 Smoothed Function	107
4.3	Unsmoothed Function versus C = 6 Smoothed Function	108
4.4	Unsmoothed Function versus Beta4 Smoothed Function	108
4.5	Unsmoothed Function versus S = 0.10 Smoothed Function	109
4.6	Unsmoothed Function versus S = 0.60 Smoothed Function	109
4.7	Unsmoothed Function versus S = 1.00 Smoothed Function	110
4.8	Smoothed Distributions for versus Log-linear (C=4), Beta4, and Cubic Spline (S = 0.60)	110
4.9	Relative Frequency Distributions for 2004 and 2003 Tests	111
4.10	Identity versus Equipercentile Equating Relationship	111
4.11	Smoothed Random Groups and External Anchor Test Equipercentile Equating Functions of 2004 to 2003 Forms	112
4.12	Smoothed Random Groups and External Anchor Test Equipercentile Equating Function	112
4.13	Smoothed Random Groups and External Anchor Test Equipercentile Equating Function	113

CHAPTER 1

INTRODUCTION

It seems quite natural for people to make direct comparisons of the educational performance of students from one year to another. Throughout the world, scores on examinations, which are regarded as indicators of students' educational performance, are compared across cohorts, and across grade levels to establish if, in fact, the system is achieving its goals. In Malawi, for example, when the public examination results have been released, there is always a debate within the media about whether educational standards are changing. The percent of students who passed that year is often compared to the previous passing percentages and a determination is made based on the magnitude of the numbers regarding whether students are doing better or worse than previous cohorts and whether educational standards are rising. Sometimes heads roll and costly reforms are initiated when standards are judged to be on the decline. It is not surprising, therefore, that examinations have been highly politicized, with the public blaming government and examination officials of professional conspiracy to fix examination results for the purposes of masking the poor performance of students. However, oftentimes the debate is short of evidence to prove the point.

The fact that such high stakes decisions are made based on test scores calls for a more serious scrutiny of the validity of comparing scores from different tests. Before making any comparison, it is important to establish how scores from different annual administrations of the test relate to each other. Unlike in physical sciences where we can directly measure whether the population is getting taller or heavier over time, the instruments (tests) used to measure educational and psychological characteristics need to

be defined within an existing context. Therefore, it becomes meaningless to directly compare scores from different contexts, unless a relationship between such scores has been defined. The statistical process for establishing this kind of relationship is called equating.

1.1 Definition and Conditions for Equating

Crocker and Algina (1986) defined equating as the process of establishing equivalent scores on two instruments. In the Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999) equating is defined as a process of placing "scores on two or more essentially parallel tests on a common scale" (p. 175). Equating, therefore, is one of the linking methods for score scale conversions aimed at achieving equivalency (comparability) of scores on two or more tests. Through equating, scores on one test are statistically adjusted for difficulty to the level of scores on another test (van Davier, Holland, & Thayer, 2004). It turns out, however, that not all tests can be equated.

There are conditions that a particular statistical adjustment must satisfy for it to be regarded as equating. First proposed by Lord (1980), these conditions are the equal construct, equal reliability, population invariance, equity, and symmetry. The equal construct, equal reliability, and population invariance conditions are satisfied when tests to be equated measure the same construct with equal reliability and are related in the same way across different subpopulations. The equity condition holds if, after equating, scores on the tests can be used interchangeably such that it should not matter which test an examinee chooses to take. This is possible only when, for every ability level, the

conditional frequency distribution of adjusted scores is the same as the conditional distribution of the original scores. Given two tests, old and new, the condition of symmetry holds if the equating function for transforming scores on the new test to scores on the old test is an inverse of the function for transforming scores on the old test to scores on the new test.

Equating is expected to satisfy all these conditions; otherwise the adjusted scores would not be interchangeable (Dorans & Holland, 2000). Consequently, equating is often considered as the most stringent of the processes for creating comparable scores. There are other weak processes that are also used to convert scores on one test to the scale of scores on another test, which are not required to meet all the conditions mentioned in this section. These processes are generally referred to as linking methods and they include: calibration, concordance, statistical moderation, and prediction. Angoff (1971), Dorans (2004), Kolen and Brennan (2004), Linn (1993), and van Davier, et al. (2004) provided excellent distinction of these methods. Note that all procedures for linking scores lead to comparable scores, but only equating provides interchangeable scores. The term 'equating' is, therefore, strictly reserved for score conversions for alternate forms of a test (i.e., tests that measure the same content and are built to the same specification) leading to interchangeable scores.

1.2 Rationale for Equating

The rationale for equating is straightforward. When two tests, old and new, have been given to different groups of examinees and scores on the new test are higher than scores on the old test, there are many ways of explaining such a difference. It could be that the new test is easier than the old one or examinees that took the new test were

brighter than those who took the old test. The need for equating arises in such situations. Since equating statistically adjusts for difficulty scores obtained on different test forms so that they are equivalent (Angoff, 1971; Kolen & Brennan, 1995 & 2004; Dorans & Holland, 2000; and van Davier et al. 2004), the relative position of examinees does not change and it becomes easier to attribute the differences in score distribution to ability differences of the groups of examinees. This allows us to determine whether one cohort performed higher or lower on the test than the other cohorts.

Equating is also instrumental in maintaining examination standards across test forms. In most testing companies, standards on examinations are usually set by experts through a well controlled process. Oftentimes, performance descriptors are used to characterize the behavior of a borderline examinee on a test and cut scores are usually set through such a characterization. In fact, this is the way cut scores are oftentimes given meaning by basing them on judgments about the adequacy of test performance (i.e., on performance levels). Unfortunately, it is usually impractical to set standards on every test form. Therefore, test versions are equated so that the ability level associated with a cut off point set on one test form remain constant over the subsequent administrations.

1.3 Equating MANEB Tests

1.3.1 Is it Necessary to Equate MANEB Tests?

It is customary for testing agencies in the United States, Canada, and some countries in Europe to equate tests as long as they are alternate forms of the same test. They usually do this even without collecting evidence as to whether it is necessary or appropriate to equate them because they have learned over the years that test forms can rarely (or if ever) be precisely equivalent in level and range of difficulty. However, such

kind of evidence ought to be the basis for the process. Equating becomes inappropriate when distributions of scores on the two forms are very dissimilar and it becomes unnecessary when distributions of scores are very similar (Harris & Crouse, 1993). Therefore, equating should only be conducted after collecting evidence that it is an appropriate or necessary process.

In this study, one hypothesis was that it is necessary to equate high stakes test forms developed by the Malawi National Examinations Board (MANEB). Because of security concerns, the Board does not reuse its items and this implies creating new tests every year. Despite the best effort by experts to match the content and difficulty levels of the forms across occasions, these test versions are dissimilar in difficulty. Therefore, fairness demands that the difficulty level of each new test form be adjusted to the level of difficulty of the old form before comparing students' performance. This hypothesis needs to be empirically tested to justify the appropriateness of equating.

1.3.2 Scaling Educational Tests

Most testing agencies transform scores in some way before reporting them. For example, prior to reporting, MANEB transforms raw scores to a 9-point scale for the Malawi School Certificate of Education (MSCE) Examinations and to a letter grade scale (A – F) for the Primary School Leaving Certificate of Education (PSLCE) Examination and Junior Certificate (JC) Examination. However, these kinds of transformation are often regarded as scaling rather than equating. Scaling serves a purpose that is different from the purpose of equating. Nevertheless, the board seems to be contented with it just like many other examinations boards in Africa and in the British examination system upon which it is modeled.

A brief survey was conducted to understand how other examinations boards in the United Kingdom and in a few former British Colonies in Africa ensure the equivalency of scores on different test forms. In response to the questionnaire, some examinations boards in the United Kingdom showed that they ensure comparability of results across tests through scaling, as oppose to equating. For example, the Assessment and Qualification Alliance (AQA) transforms raw scores to a Uniform Mark Scale (UMS) for all General Certificate of Education (GCE), Vocational Certificate of Education (VCE), and modular and non-modular General Certificate of Secondary Education (GCSE) examinations (AQA Uniform Marks Leaflet, 2004). Similarly, the Welsh Joint Education Committee (WJEC) scales results on their GCE and GCSE examinations to a letter grade scale (G. Kelly, personal communication, February 17, 2006). This later example compares favorably with the way MANEB transforms its scores. The National Examinations Council of Tanzania (NECTA) and the Independent Examinations Board (IEB) of South Africa also transform scores on test forms in the same way. While scaling test scores is an important process, it cannot replace equating.

Scaling places scores on different tests or test forms on the same scale. Since it is a linear transformation scaling preserves the rank ordering of students, but it does not adjust scores for difficulty of the test forms. For example, a student classified as failing on the raw score scale will remain a failing student after scaling. Therefore, fair comparison of scores across years still requires a different kind of transformation. Equating is well placed to convert scores for purposes of generating equivalent scores across test forms. In cases where scores are reported on the same scale every year, equating should precede scaling.

1.3.3 Standards of Examinations

One of the desirable goals for many assessment systems is meeting the requirement that standards of examinations be maintained over years. This, at least, seems to be the expectation of consumers of education in Malawi. The public believes that the passing mark on all MSCE exams is maintained at 33%. In fact, the Parliamentary Committee on Education (PCE) is currently investigating the allegation that MANEB lowered the passing score on the 2005 examinations from 33% to 16% because of political pressure to pass many candidates. Interestingly, while the board denied that the passing score has never been fixed at 33%, it failed to tell the public the correct passing score. While these allegations may be untrue, they reflect the existence of mistrust between the examinations boards on one hand and the public on the other. It is easy to notice that the public has a belief that appears to run counter to what actually happens. However, the public will continue making unsubstantiated allegations as long as they remain uninformed and as long as the testing industry does not make the process of setting and maintaining standards as transparent as possible. The onus, therefore, is on the board and researchers to explore sound psychometric practices to meet the expectations of the society without jeopardizing the integrity and security of the exams.

1.3.4 Lack of Appropriate Equating Design

There are reasons why examinations boards like MANEB may not equate its tests despite the willingness to carry out the process. One of them is lack of a suitable equating design. A discussion of the designs used in equating has been presented in the literature chapter of this dissertation. But suffice it to say that there are three popular equating designs: single group, random groups, and the non-equivalent groups with anchor test

(NEAT) designs. For testing agencies that administer high stakes achievement tests on a yearly basis, the single group and the random groups designs are not suitable because the populations that take each test are different and the second form is often administered when the other group is out of school or grade.

The only suitable design is the NEAT design where a set of common items is included in each test. These anchor items are used to determine differences in ability of the two groups and to provide a basis for disentangling test and group differences. However, for anchor items to work properly they are not released. The behavior of disclosed items does not remain invariant since students use them for practice, which will make them look easier during second administration. This requirement presents a special challenge to testing agencies like MANEB that releases all the items after each administration. Therefore, for MANEB, the way forward is to come up with its own design or modify the already existing ones to suit its situation.

1.4 Statement of the Problem

Research has not explicitly shown, using empirical data, the consequences of not equating educational tests across years or occasions. As such, the public, the media and some educators in countries like Malawi continue to make direct comparison of the scores across cohorts to establish if the educational system is achieving its goals. Worse still, there has been no research to explicitly show that decisions made based on scores that are not equated may be flawed. There is also no empirical evidence to support the assumption that the examination standards for such tests have remained invariant over the years. Therefore, for high stakes examinations such as the MANEB's tests, these are relevant issues that warrant further investigation.

Examinations boards that do not equate tests certainly do not see the appropriateness of the process. For Malawi, this seems to be a practice that was inherited from the British examination systems and it is still perpetuated today. Test scores across years are considered comparable as long as the forms that are used to generate them are supposedly measuring the same construct with the same specifications. They are contented with scaling and the use of the same benchmarks across forms. In the absence of empirical justification for the practice of equating, there is nothing to argue against such people. Therefore, a study that makes a case for equating is warranted.

Assuming there is an interest to equate tests, there seems to be no obvious equating design that is appropriate for boards like MANEB to use. This is because test items are disclosed after each administration. As noted earlier, these exposed items cannot be reused as anchor items in the next administration. Other equating designs like random groups or single group designs are not suitable for the board since examinations are administered once every year to different populations of students. If it is to equate its tests, MANEB has to create its own design or modify the already existing designs to adapt them to its situation. Therefore, research is needed to understand which design may be appropriate.

1.5 Purpose of the Study

There were three purposes in this study. The first was to collect evidence regarding whether or not it is necessary to equate test forms developed by MANEB. On this point the focus was to investigate the degree of similarity in test difficulty and in score distributions across test forms. The evidence was intended to support the decision for or against equating.

The second purpose was to investigate the consequences of not equating educational tests. The consequences of interest were the invariance of the examination standards (cut scores) across years, and the effect of using scores that are not equated to classify students. This investigation, too, was intended to support the case for equating.

The third purpose was to propose an equating design that may be used in situations where all the items are disclosed. On this aspect, the interest was to investigate the possibility of using an external anchor test that is administered separately from the operational tests to equate scores. This is an adaptation of the non-equivalent groups with anchor test (NEAT) design. Because of the time constraints, the design was evaluated using the random (equivalent) groups design.

The rest of this document will proceed by providing a review of procedures for assessing whether it is necessary to equate two test forms, followed by a review of the major equating designs and methods. Attention is drawn to the applicability of these designs to the situation where test forms are administered once a year to a different population of examinees and to a situation where all items are exposed. Lastly, previous attempts to equate educational tests using external anchor tests will be detailed.

CHAPTER 2

LITERATURE REVIEW

This chapter begins with a description of procedures used in investigating whether or not equating is necessary. It continues by describing steps involved in the equating process followed by a discussion of equating designs, and equating methods while highlighting the actual steps, designs, and methods to be implemented in this study. Next, a review of earlier attempts in equating scores using an external anchor test is given. In that section, different kinds of anchor tests used by researchers over the years have been discussed to highlight the different ways of constructing and administering anchor tests. Following that, a description of procedures for evaluating the adequacy of equating is provided. The discussion is intended to form a basis for the procedures used in this study. The chapter ends with a summary of the reviewed literature and an explanation of the knowledge gap which this study intends to fill.

2.1 Determining Whether to Equate

The study, among others, intends to make a case for equating by collecting evidence to support the idea that it is necessary to equate educational tests. To accomplish this task, it is important, during the investigation, to employ procedures that are well grounded in literature. This section discusses major procedures that researchers have used to determine the whether to go ahead with equating. According to Harris and Crouse (1993) equating may sometimes be inappropriate or unnecessary. It becomes inappropriate when data from the tests to be equated are very dissimilar and it becomes unnecessary when data from the tests to be equated are very similar. This section reviews literature on these two categories of ideas.

2.1.1 Is Equating Appropriate?

It is well known that not all tests should be equated (Angoff, 1971; Dorans, 2004; & Kolen, 2004; Lord, 1980). In some cases, equating is capable of adding in more error than it may remove (Harris & Crouse, 1993). Therefore, the decision to equate tests should be based on a priori evaluation that would help support the decision to equate test scores. There are many procedures found in literature for investigating the appropriateness of equating in a particular situation.

Loret (1972, cited in Harris & Crouse, 1993) equated seven standardized tests that were not designed as parallel forms. The disattenuated inter-correlations of the tests were used to determine the appropriateness of equating. With this procedure, tests should have a disattenuated correlation of .95 or higher to proceed with equating. This arbitrary, but acceptable criterion is not popular in literature because two tests measuring entirely different things may correlate highly in a particular sample of examinees (Harris & Crouse, 1993) and it requires two test administrations which may not be practically feasible for most testing agencies.

Kolen (2004) proposed an evaluation framework for determining the degree to which equating, calibration, and concordance can be achieved given a particular situation. The proposed framework is as follows: (1) Inferences - to what extent are scores for the two tests used to draw similar inferences? (2) Constructs - to what extent are the two tests measuring the same construct? (3) Population - to what extent are the two tests designed to be used within the same population? (4) Measurement conditions - to what extent do the tests share common measurement conditions, including, for example, test length, test format, administration conditions, and so on? Based on this framework, tests that

measure different constructs in different populations, tests whose scores are used to draw different inferences, or tests that have different measurement conditions cannot be equated. Both judgmental and statistical processes are required to carry out such evaluation. Although this seems to be adequate criterion, Kolen (2004) still called for the development of "systematic judgmental (and statistical) procedures for analyzing the similarity of tests and testing conditions to assess whether equating or concordance between two tests is likely to be possible and useful" (p. 225).

Another important contribution to the development of a priori evaluation framework was made by Dorans (2004) who suggested procedures for judging the similarity of the tests. These procedures would also help determine the extent to which equating, concordance or prediction can be achieved. The first criterion is to evaluate the similarities in constructs that are measured by the tests, which is accomplished by evaluating the similarities in content and in test specifications through judgmental means. The second criterion is to evaluate the strength of the empirical relationship between the scores that are to be linked through analyses such as factor analysis, structural equation modeling and correlations. The third criterion is to assess the extent to which equating is invariant across subpopulations. In this context, the degree to which the same equating function can be derived from different subpopulations (e.g. boys and girls) is evaluated.

The present study intends to equate the 2003, 2004, and 2005 PLSCE math tests developed by MANEB. The tests meet the criteria proposed by Kolen (2004) in that they are alternate forms (of course, not strictly parallel) designed to measure the same construct, mathematics proficiency; the results of these tests are often used to make similar inferences about math proficiency of the examinees; they are all designed to be

used in grade eight in the country; and finally the tests have same item format, equal test length and they are administered under similar conditions in all primary schools. Of course, the structures of the test forms were not evaluated to establish if they meet the Dorans's criterion. Nevertheless, equating such tests seems appropriate.

2.1.2 Is Equating Necessary?

Using the same reasoning by Harris and Crouse (1993) that poor equating is capable of adding in more error than it may remove, it becomes unnecessary to equate tests if their score distributions were very similar. The critics of equating may certainly embrace this idea, which is legitimate. However, it has to be established and not just assumed or claimed. There are also procedures that can be used to make such an evaluation.

Kolen and Brennan (2004) discussed the use of identity method to show that sometimes equating can be deemed unnecessary. They bring in the term "identity equating" to refer to a kind of conversion where "a score on Form X is considered to be equivalent to the identical score on Form Y" (p.34). They noted that identity equating would be the same as mean and linear equating if the two forms were identical in difficulty all along the score scale. Therefore, to determine whether or not to proceed with equating, one would compare the mean, linear, or equipercentile equating lines against the identity equating line. Kolen and Brennan (2004) offered guidelines for the use of identity equating. They recommended identity equating in situations where there are: poor quality control conditions, very small samples, similar test form difficulty, where simplicity is desirable for purposes of easy communication to non-psychometricians, and where inaccurate results can be tolerated.

Dorans and Lawrence (1990) used the identity method to determine if it was necessary to equate scrambled test versions to the base test version for operational use. Whenever the identity line fell within reasonable confidence interval (± 2 standard errors) after equating, then they considered equating unnecessary.

Hanson (1992, cited in Kolen & Brennan, 2004) proposed the use of log linear method to investigate if the distributions of any two tests were similar enough to warrant equating unnecessary. Using this procedure, whenever the chi-square test of the null hypothesis that the distribution of raw scores on tests is the same is rejected, then equating is necessary. If the null hypothesis were not rejected, then equating is unnecessary and instead identity equating is considered. This method seems to be less subjective (Harris & Crouse, 1993; Kolen & Brennan, 2004) than the identity equating method and it is particularly useful when the sample size is small.

In this study, both the identity and log-linear methods have been used to determine whether or not equating is necessary. It was important to establish the necessity of equating to avoid adding in more error than the amount that the process intends to remove.

2.2 Steps for Implementing Equating

The early references to test equating, which illustrated the need for and ways of obtaining comparable scores included Otis (1922), Thorndike (1922), and Kelley (1923). Several other publications addressing specific equating practices have been made since then. Many issues, however, that are involved in actually doing the test equating are discussed by Kolen and Brennan (1995 & 2004), and van Davier et al. (2004). Note that these steps are not necessarily mutually exhaustive.

Harris and Crouse (1993) identified three steps often spelled out in literature that need to be considered when conducting equating namely: (a) selecting a data collection design, (b) selecting an operational definition of equating (often a choice between linear and curvilinear methods), (c) selecting the particular estimation method to achieve the second step (such as deciding between the Tucker and Levine linear methods), and (d) selecting evaluation criteria. The choice of data collection design, determines the definition of the equating relationship to be used and the estimation methods to be applied. They added a fourth step to help determine if the method chosen results in an equating of adequate accuracy.

Kolen and Brennan (2004) presented seven steps that may be followed when equating tests: (1) Decide on the purpose for equating; (2) Construct alternate forms; (3) Choose a design for data collection; (4) Implement the data collection design; (5) Choose one or more operational definitions of equating; (6) Choose one or more statistical estimation methods; and (7) Evaluate the results of equating. They argue that these steps are important to successful equating. It is always important to have a reason for conducting equating and the first step satisfies this requirement. Since only tests that are parallel in content and statistical specifications (i.e., alternate form) are equated, the second step aims at creating these tests. For this study, tests to be equated were already constructed by MANEB and they were re-administered to collect data that intended to address the research questions. Other than the first two, the Harris's and Crouse's (1993) steps are essentially similar to the Kolen's and Brennan's (2004) proposed steps. The present study follows steps 1, 3, 4, 5, 6, and 7. The equating designs, equating methods, and evaluation criteria are the subjects of the rest of this chapter.

2.3 Equating Designs

There are several data collection designs (equating designs) used in test equating. A complete description of these designs is given by Angoff (1971), Holland and Rubin (1982); Kolen and Brennan (1995 & 2004), and van Davier et al. (2004). The choice of which designs to use depends on a host of factors such as test security issues, availability of the sample, time frame, and many others. In any case, van Davier, et al. (2004) argued that equating almost always seeks to control for differential examinees' ability as it controls for differential difficulty of the tests. It turns out that employing designs that use equivalent groups or common items can control for differential examinees' ability. Van Davier, et al., (2004) categorized the designs as follows: Those that use equivalent groups are Single Group (SG), Random Groups (RG), and Counterbalanced (CB) designs and the design that uses common items is the Non-Equivalent Groups with Anchor Test (NEAT) design. What follows is a brief description of these designs. The section has also highlighted the design used in this study and factors that dictated the choice.

2.3.1 Single Group (SG) Design

The single group (SG) design controls for differences in examinees' ability by having the same group of examinees take both tests. For this design, van Davier, et al., (2004) explained that any differences in the scores and in the score distributions are attributed to differences in test difficulty because the tests are administered to the same students. However, Kolen and Brennan (2004) noted a number of shortfalls for this design. Among them, the design is affected by order effects, practice effects, and fatigue. Since it requires two administrations of the test, it implies doubling the testing time. Oftentimes, testing companies do not have such luxurious time. In fact, oftentimes tests

to be equated are administered to different groups of examinees in different years or occasions. Therefore, this design is not appropriate for most of testing organizations including MANEB and as such is rarely employed. Nevertheless, the SG design may be used to collect data during filed testing or during research studies.

2.3.2 Counterbalanced (CB) Design

Kolen and Brennan (2004) regard this design as part of the single group design. They refer to it as the single group design with counterbalancing. van Davier, et al., (2004) regard it as a separate design with assumptions of the SG and RG designs. Nevertheless, the counterbalance (CB) design is a variation of the SG design, which also controls for differences in examinees' ability using equivalent groups. In this design, tests to be equated are administered to two random samples of examinees from a single population in different order. One group takes the new test first and the old test second, whereas the other group takes the old test first and the new test second. Therefore, the CB design controls for differential order effect and fatigue, which restrict the usefulness of the SG design. However, like the SG design, the administration of two tests to the same group of students is still not practically feasible for most testing agencies.

2.3.3 Random Groups (RG) Design

van Davier, et al. (2004) refers to the random groups (RG) design as the Equivalent Groups (EG) Designs. It controls for differences in examinees' ability by drawing two randomly equivalent samples of examinees from a common population and one group takes the new test whereas the other group takes the old test. Kolen and Brennan (2004) recommended the use of spiraling procedures to create these equivalent groups. If the samples are large enough, differences in score distributions are attributed to

differences in test difficulty since samples are considered equivalent. The RG design is an improvement over the SG design because it requires the administration of only one test per group. Therefore, it is more practical and the problem of practice effect due to familiarity, and fatigue are solved. However, like the SG design, it cannot be used to equate scores on tests that are administered in different years. Furthermore, when the samples are small, errors can be very large (Kolen & Brennan, 2004).

2.3.4 Non-Equivalent Groups with Anchor Test (NEAT) Design

For many large-scale testing programs, the Non-Equivalent Groups with Anchor Test (NEAT) design is preferred because of its administrative flexibility, allowing only one test form to be administered to groups of examinees that are not necessarily equivalent. Since it allows groups to be non-equivalent, it can be used to equate tests given in different years to different groups of people provided the two groups also take a set of anchor items. The anchor items are used to determine differences in ability of the two groups, which in turn provides a basis for adjusting scores on the tests. However, strong statistical assumptions are usually made to disconfound group and test differences (Kolen & Brennan, 2004). Anchor items can be internal or external (Lord, 1980; Kolen & Brennan, 2004). They are regarded as internal when a score on these anchor items contributes to an examinee total score on the test. In an external anchor design, the items do not contribute to the examinee's final score on the test.

In this study, the RG design was used to collect data for equating the 2005, 2004, and 2003 test forms. There were a number of factors that dictated our choice. One factor was that, at the time of the study, the populations of students that took the 2004 and 2003 had graduated and there were no data for these forms. Even if the data were available,

there were no common items on the tests such that the NEAT design could not be used. The only way to collect data for the study was to re-administer the forms to the 2005 population. The SG could not be used because it requires administering all the test forms to the same group. These circumstances necessitated the use of the random groups design. Although the study equated tests from different years, the RG was appropriate because the goal was to make a case for equating by investigating important concepts related to the process. Therefore, it was purely for illustrative purposes. Kolen and Brennan (2004) qualified the RG design as "...ideal for presenting many of the statistical concepts in observed score equating" (p.26) because, comparatively, it requires very few statistical assumptions, which are most readily achieved.

The study also used a modified NEAT design to investigate the possibility of using an external anchor test to equate test forms. Explanations qualifying the difference between what is known and what is not known about this design are given later in the chapter.

2.4 Equating Methods

Equating methods refer to a collection of techniques that have been developed to solve the score equating problems that have arisen in a wide variety of practical testing circumstances (Dorans, 2004). Dorans and Holland (2000) categorized them into two: those that use observed scores and those that make use of 'true scores.' This study focuses on the observed score equating. Several of these methods (procedures) that were developed between 1920 and 1970 are described in detail by Flanagan (1951), Gulliksen (1950), and by Angoff (1971). Lord (1980), Braun and Holland (1982), and Morris (1982) provided a mathematical treatment for test equating. More recent authorities,

however, include Kolen and Brennan (2004), Livingston (2004), and van Davier, et al. (2004). However, there are many other publications that have described and compared the performance of the equating methods in different situations (Budescu, 1987; Dorans, 1990, Harris & Kolen, 1990).

Kolen and Brennan (2004), and Livingston (2004) presented three methods of observed score equating, which are discussed in this section: (1) Mean Equating; (2) Linear (mean and sigma) Equating; and (3) Equipercentile Equating. These methods differ in the way each one of them defines equivalent scores. Livingston (2004) noted that the different definitions of equivalent scores arise from the use of different definitions of the 'relative position of scores' in a group of examinees. Anyone of these equating methods can be used to equate test scores under each of the equating designs discussed in the preceding section. However, since different equating designs offer different information, and since the designs make different assumptions about the groups of examinees, the equating methods use the information differently. For example, a linear equating relationship for the RG design is defined differently from the linear equating relationship for the NEAT design. The definitions of the equating methods presented in the following sections, relate to the random groups (RG) design because, it is this design that was used to collect data for the study.

To enhance understanding of the mathematical expressions presented in this chapter, symbols proposed by Kolen and Brennan (2004) have been adopted. The symbols are defined as follows: The new test form has been defined as Form X whereas the old form has been defined as Form Y. X is the random variable scores on Form X, with x as a particular score on Form X (i.e., a realization of X). Similarly, Y is the

random variable score on Form Y, and y is the realization of Y. These symbols are used to mathematically describe the mean, linear, and equipercentile equating methods under the random groups design. Note that the descriptions of the methods presented in this chapter may be different for other equating designs.

2.4.1 Mean Equating

Mean equating defines relative position in terms of the number of points above or below the mean in the target population of examinees (Livingston, 2004). Therefore, in mean equating, equivalent scores are obtained by setting equal scores on the two test forms that are equal (assigned) distance away from their respective means as shown in the mathematical expression below:

$$x - \mu(X) = y - \mu(Y)$$

In this expression, $\mu(X)$ and $\mu(Y)$ are the means of X and Y respectively. Thus, this method of equating is simply implemented by setting the deviation scores on the two test forms equal. The conversion, $m_y(x)$ for transforming a score on Form X to the scale of scores on Form Y, is obtained by solving the expression for y:

$$m_y(x) = y = x - \mu(X) + \mu(Y)$$

In this way, the examinee's adjusted score will have the same relative position (number of points above or below the mean) in the target population as her score on the new test has in the target population. The $m_y(x)$ is obtained by adding a constant to all raw scores on Form X to find equated scores on Form Y. Mean equating is rarely done in practice because the new form is always considered to differ from the old form by a constant. In this way, it does not take into account how high or low an examinee's score is in the distribution of scores. However, it may be used to illustrate an important concept.

2.4.2 Linear Equating

The most familiar and widely used of all equating methods is the linear equating (van Davier et al. 2004). In contrast to mean equating, linear equating defines relative position in terms of both the mean and standard deviation. This kind of adjustment takes into account how high or low the examinee's score is in the distribution of scores. That is, it allows for the test forms to be differentially difficult along the score scale (Kolen & Brennan, 2004). Equivalent scores are obtained by transforming scores on the new form to scores on the old form that are the same number of standard deviations above or below the mean of the group (Livingston, 2004). That is, setting the standardized deviation scores (z-scores) on the two tests to be equal as shown here:

$$\frac{x-\mu(X)}{\sigma(X)}=\frac{y-\mu(Y)}{\sigma(Y)}$$

where $\sigma(X)$ and $\sigma(Y)$ are the standard deviations of X and Y respectively. As previously defined, $\mu(X)$ and $\mu(Y)$ are, respectively, the means of X and Y. The conversion, $l_y(x)$, for transforming a score on Form X to the scale of Form Y is obtained by solving the expression for y and rearranging terms to get:

$$l_y(x)=y=\frac{\sigma(Y)}{\sigma(X)}x+\left[\mu(Y)-\frac{\sigma(Y)}{\sigma(X)}\mu(X)\right]$$

Using this linear conversion, the adjusted scores on the new form will have the same mean and standard deviation as the raw scores on the old test. Linear equating is based on the assumption that the distributions of X and Y differ only in the means and standard deviations (Crocker, & Algina, 1986). Livingston (2004) outlined the following downsides to linear equating: A very high score or low score on the new form can equate to score outside the range of possible scores on the test; the results of linear equating

depend heavily on the group of examinees; and when the two tests differ in difficulty, linear equating in a strong group of examinees will differ noticeably from the linear equating in a weak group of examinees (i.e., population dependence). The problem of out-of-range is sometimes solved by truncation, which involves setting adjusted scores that exceed than 100 equal to 100 and all adjusted scores lower than 0 are set to 0.

2.4.3 Equipercentile Equating

The equipercentile equating method minimizes the out-of-range problem and it takes into account the possibility that the target population's score distributions on the new test and the old test may have different shapes (Livingston, 2004). The method defines relative position in terms of percentile ranks. It considers a score on the new test to be equivalent to a score on the old test if they have the same percentile ranks in the population of examinees. Therefore, to carry out equipercentile equating, scores on the new form are transformed to scores on the old form that have the same percentile rank in the target population. This kind of conversion makes the distribution function of scores on Form X converted to Form Y scale equal to the distribution of scores on Form Y in the population (Kolen & Brennan, 2004).

In mathematical terms, if F is the cumulative distribution function of X in the population, G is the cumulative distribution function of Y in the same population, e_y is a symmetric function used to transform scores on Form X to the scale of Form Y, and G^* is a cumulative distribution function of e_y in the same population, then e_y is defined to be an equipercentile equating function in the population if: $G^* = G$. Braun and Holland (1982), Kolen and Brennan (2004) and van Davier et al. (2004) regarded the following as an eqquipercentile equating function:

$$e_Y(x) = G^{-1}\left[F(x)\right]$$

where G^{-1} is the inverse of the cumulative distribution function of G. The main problem with equipercentile equating is that the score distributions on real tests are often irregular. Thus the percentage of examinees with a given score fluctuates as the scores increase. These fluctuations in turn produce irregularities in the equipercentile equating adjustment, which do not generalize to other groups of examinees. The problem, however, is sometimes solved by incorporating smoothing procedures in the equating process to remove the irregularities.

There are two types of smoothing methods, pre-smoothing and post-smoothing. In pre-smoothing, the score distributions of the test forms to be equated are smoothed whereas in post-smoothing, the equipercentile equivalents are smoothed. The principle behind smoothing (both pre-smoothing and post-smoothing) is to remove the irregularities while preserving the location, spread, and shape of the score distribution (Kolen & Brennan, 1995 & 2004; Livingston, 2004; and van Davier et al., 2004). When smoothing is employed, total equating error is partitioned into random error and systematic error. The random error is what makes the score distributions irregular and it is this random error that smoothing tries to reduce. The systematic error is introduced, among other, by the smoothing process. Smoothing methods tries to produce smooth functions with less random error than that for unsmoothed distributions. This is done by increasing systematic error in such a way that it becomes more than offset by the decrease in random error. Therefore, smoothing is successful to the extent that it results in less total equating error than that for the unsmoothed distributions regardless of the systematic error it introduces into the distributions.

2.4.4 Other Equating Methods

Information from anchor items for the NEAT design is used in many ways leading to different anchor test equating methods. Kolen and Brennan (2004), Livingston (2004), and van Davier, et al. (2004) identified two ways in which information from anchor items is used: Chain Equating and Post-Stratification Equating. In chain equating scores on the new test are equated to scores on the anchor and scores on the anchor are equated to scores on the old test. The scores can be equated using linear (chain linear equating) or equipercentile (chain equipercentile equating, or Lindquist equating) methods. Dorans and Holland (2000) presented the assumptions for these kinds of equating. They noted that chain equating assumes the equating relationship used to equate scores on the new test to scores on the anchor is population invariant (i.e., it generalizes from the equating sample to the target population) and the function used to equate scores on the anchor to scores on the old test is also population invariant.

Livingston (2004) offered a simple explanation for post-stratification equating. In this particular kind of equating, the set of anchor items is used as if it were a predictor variable. For every score on the anchor, marginal distributions of X and Y are estimated. The estimates are then used in equating as if they had actually been observed in the population. According to van Davier, et. al. (2004), to estimate the distributions of X and Y, post-stratification equating methods assume these distributions are conditional distributions and that the conditional distribution of X given the anchor test, and conditional distribution of Y given the anchor test, are population invariant (i.e., they generalize from each sample to the target population). The post-stratification equating methods can be linear or equipercentile methods (Livingston, 2004). Linear methods

include the Tucker and Levine equating methods whereas the most notable equipercentile method is the frequency estimation method. Detailed description of these method are presented by Dorans and Holland (2000), and Kolen and Brennan (1995 & 2004).

In this study equipercentile equating was the method of choice for equating scores on the tests for three reasons: (1) It is based on a better definition of "relative position" of a particular score in the distribution of scores than linear and mean equating; (2) It takes into account the possibility that the target population's score distributions on the new form and on the old form may have different shapes; and (3) It minimizes the problem of out-of-range adjusted scores. However, linear equating was also used for purposes of comparing the results. For the data collected through the external anchor test design, the study used frequency estimation method, an equipercentile procedure belonging to the post-stratification equating method. The choice of this method over the chain equating was arbitrary. However, its choice over the Tucker and Levine was dictated by the need to compare equated scores from the random groups design with the equated scores from the external anchor test design. It was important to use similar (equipercentile) methods in the two equating processes to facilitate comparison. Nevertheless, Tucker method was also used to provide alternative conversion tables.

2.5 Equating Using Anchor Test

Many studies have looked at equating using anchor tests. The studies may be classified into three categories depending on the nature of the anchor test investigated. They include studies that investigated the use of: (1) common items as anchor, (2) other tests as anchor, and (2) variables as anchor. This section discusses only a few selected studies in each category to highlight these types off anchor tests.

2.5.1 Using Common Items as Anchor

When common items constitute an anchor test, the resulting anchor is either internal or external (Lord, 1980; Kolen & Brennan, 2004). A score on the internal anchor test contributes to an examinee's total score on the test whereas a score on the external anchor test does not contribute to the examinee's final score on the test. Furthermore, there are three important points about this type of anchor tests to be highlighted: The common items are oftentimes drawn from the reference form; the anchor tests tend to be shorter in length than the test forms to be equated; and the common items and the test forms to be equated are usually administered concurrently. The internal and external anchor tests are usually discussed together because there are few, if any, differences between them and they are affected by similar factors. Many researchers have studied the usefulness of common items as anchor and this section only looks at a few of them.

Angoff (1982) investigated the use of an external anchor test to equate the then Scholastic Aptitude Test (SAT) -Verbal to itself. The goal of the study was to investigate the effectiveness of both internal and external anchor tests. Conditions such as location, difficulty and content of the anchor in relation to the total test were manipulated one at a time. The equatings of the SAT-Verbal to itself through the anchor test were carried out for random samples, similar samples and dissimilar samples. The anchor test was a non-operational section of the verbal material given with the SAT and as such it was similar in content to SAT-Verbal test. The anchor contained all the four types of items found in the operational verbal test and their average difficulty was about half of the average difficulty for the SAT-Verbal. The scores were equated using the different equating methods described and results were compared to each other.

The study provided many interesting results, but of particular interest in this review are those related to the anchor test. In general, the anchor tests with similar content yielded satisfactory results when equipercentile equating methods were used in random groups design. However, anchor tests with dissimilar content did not yield satisfactory results. Therefore, internal and external anchor tests work well when they are similar in content and difficulty to the test forms to be equated.

Klein and Jarjoura (1985) also studied the importance of content representation for common item equating with non-equivalent groups of examinees. They concluded that if content representation is different between the set of anchor items and the test forms to be equated, the anchor items will not accurately reflect group differences in ability. Petersen et. al. (1983), and Wingersky et. al. (1987) studied the characteristics of common items and they reported that using large numbers of anchor items reduces the amount of random error. Angoff (1971) and Kolen and Brennan (1995 & 2004) establish a rule of thumb that the length for a set of common items should be at least 20% of the length of the total test.

The similarity of the context in which anchor items are presented in the two tests is as important as the similarity in the content. Zwick (1991) wrote about the investigations that were carried out to understand why there were large and dramatic changes in reading proficiency between 1986 and 1984 on the National Assessment of Educational Progress (NAEP) Reading. The investigations concluded that the large changes were, among others, caused by the difference in context in which the anchor items appeared in the two years, rather than the changes in the reading achievement. Anchor items in 1986 test were placed in different positions from the positions they

occupied in 1984. Kolen and Brennan (1995) noted that context effects like these can lead to very misleading results because they make anchor items behave differently in the old form and the new form. Following these studies, there are some guidelines that have been put in place to help when creating or assembling common items. Livingston (2004) provides a summary of these guidelines:

- Include enough questions from the reference form.
- Choose questions that resemble the full test in terms of content and format.
- Select questions that represent the full range of difficulty.
- All questions that have been changed should be excluded.
- Consider excluding questions at the end of the test as anchor items.
- Do not change the position of the items.
- Do not break up an item set.
- Select items that correlate well with the total score.
- Use common items that are clean – good wording and understandable.

Kolen and Brennan (1995 & 2004) recommended that anchor items should not be disclosed to ensure that they behave the same way across the different administrations. This requirement runs counter to policies in some states and countries where testing agencies are mandated to disclose items after administration for instructional purposes. For example, MANEB discloses all the items and therefore, such items cannot be reused as anchor in the next administration because the invariance of their behavior cannot be guaranteed. Although there has been little research to empirically establish the effects of using disclosed items as anchor, the use of a NEAT design with an external anchor test is usually the popular choice in such circumstances. All items that contribute to an

examinees score are disclosed for instructional purposes whereas the external anchor test is not disclosed. If items in the target and anchor tests are administered concurrently, it is difficult to disclose just the target test items and collect back the anchor items. It is for this reason that this study investigated the use of external anchor items to equate tests given in different years.

2.5.2 Using a Different Test as Anchor

Angoff (1982) conducted another study in which different tests, Test of Standard Written English (TSWE) and SAT-Mathematical, were used to equate SAT-Verbal to itself. The TSWE and SAT-Mathematical were operational tests administered with the SAT-Verbal, but they were regarded as external anchor tests. The content of each test was dissimilar to that of the SAT-Verbal since they did not include items primarily intended to measure reading skills. In terms of difficulty, the TSWE was more dissimilar to SAT-Verbal than was the SAT-Mathematical, but TSWE was more similar in content to SAT-Verbal than SAT-Mathematical. Again the equating was done using random samples, similar samples and dissimilar samples. Different equating methods as described in the previous section were used in the equating process.

The study found that the total errors of equating tended to be greater for all equating methods in the random samples design when TSWE was used as the external anchor test than when SAT-Mathematical was used as the external anchor test. This finding demonstrated that equating results may be affected more by differences in difficulty between the anchor test and the operational test than by differences in content. Another finding was that in most circumstances random samples yielded more satisfactory results than dissimilar samples. The study demonstrated that different tests

could be used as anchor when random samples are used in the equating process. However, "an external anchor test that is constructed to be a miniature of the total test is required when nonrandom samples, particularly dissimilar samples, are used" (p. 103). This finding is explains why anchor test items for use in the NEAT design are usually drawn from the reference form.

It is important to highlight a few points again regarding the different tests as anchor. The tests were as longer in length as the target test forms to be equated; and the target test and the different tests as anchor were administered together. However, the items in the anchor were not drawn from the reference form and the items in the different tests would be changed in the next administration.

2.5.3 Using Other Variables as Anchor

Scores on common items scores do not always correlate highly with scores on target tests (Wright & Dorans, 1993; Liou, Cheng & Li, 2001). Also, because target tests are frequently administered at different occasions, scores collected at the second testing occasion might be contaminated by nonrandom errors due to test disclosure (Liou, Cheng & Li, 2001). Wright and Dorans (1993) suggested using selection (surrogate) variables (e.g., school grades, other test scores) as the anchor to account for group differences. In their study, Wright and Dorans (1993) investigated whether equating results can be improved if the variable that accounts for all systematic differences between equating populations is identified and used as an anchor in anchor test design or as a variable on which to match equating samples. They used selection variables, math scaled scores and verbal scaled scores, to equate Scholastic Aptitude Test (SAT) Math forms and SAT Verbal forms respectively. The sample invariant properties for four anchor test equating

methods (Tucker, Levine, chained equipercentile, and frequency estimation equipercentile models) were examined under representative, matched-on-equating-test, and matched-on-selection-variable conditions. The selection variable, the variable along which subpopulations differ, was also used as an anchor for four equating methods and compared to equatings in which the equating test served as an anchor. All equatings were performed with real Scholastic Aptitude Test (SAT) populations or simulated populations.

The study showed that matching on the selection variable improved accuracy of equating over matching on the anchor test for all methods. Results with the selection variable as an anchor were good for both the Tucker and frequency estimation methods, but unacceptable for Levine and chained equipercentile results. In fact, "both the Tucker and frequency estimation procedures performed better, in most cases, with the selection variable as an anchor than they did when the common items served an anchor" (p. 22). The authors explained these surprising results by noting that the Tucker and frequency estimation methods assume that the anchor test they are using is, in effect, the variable along which the old and new form samples differ. Therefore, their assumptions were not violated when the selection variable was used as an anchor test. On the other hand, the use of selection variable as an anchor produced unreasonable results for Levine and chained equipercentile because their assumptions were violated. The Levine model assumes that the true score correlation between the anchor test and the test to be equated is unity, which is not usually the case in practice. The chained equipercentile performed poorly because the scaling relationship between verbal and math across populations differs systematically, which was in violation of the model's population invariance

assumption. One implication of this finding is that school variables as anchor may only be used with certain kinds of equating methods. This restriction is certainly unpleasant where many other factors dictate the choice of the method. Therefore, the use of selection variables as anchor is less appealing.

Liou, Cheng and Li (2001) studied a similar problem to that of Wright and Dorans (1993). In their study, they equated two forms of a test administered to nonequivalent groups with common items in three schools. The common items were similar in content and difficulty to the target forms. They also used examinees' average school scores in Geography and for each examinee, the Geography score served as a surrogate for the common item score. Different methods including the frequency estimation and the imputation approach using either common items or Geography scores were used to estimate comparable scores for the test forms. The results suggested that a Geography score worked as well as the common-item score, even though the two variables had lower correlations with the target tests. Note that the variables serving as or complimenting the anchor test may have been administered separately from the target test forms and they are not drawn from any of the target tests to be equated. It is clear not whether the variables were from schools or they were part of the scores in the same testing system.

2.6 Criteria for Evaluating the Adequacy of Equating

Harris and Crouse (1993) reviewed several ways that researchers use to evaluate the adequacy of equating. They considered the following criteria: weak equity, indices (such as root mean square error), standard errors of equating, generated data, equating a test to itself (self or circular equating), large sample, consistency across methods, replication and cross validation samples, heuristic, and other methods. One concern about

these methods was that they sometimes lead to different conclusions regarding the adequacy of equating. The review in this section is only intentionally centered on a few of these criteria in order to provide the basis for methods used to evaluate the adequacy of equating in the study.

2.6.1 Standard Error of Equating

Many studies have used the standard error of equating before to estimate the degree of precision with which scores have been equated or to compare the precision of equating across methods (Angoff & Cowell, 1986; Fairbank, 1987; Jarjoura & Kolen, 1985; Kolen, 1985; Lord, 1982; Wang, et. al., 2000, Ogasawara, 2001). But what are equating errors and how are they estimated?

The equating functions estimated using any of the methods mentioned earlier are subject to sampling variability since they are estimated from the sample estimates of the population parameters (Kolen & Brennan, 1995 & 2004; van Davier, et. al., 2004). The sampling variability of the equating function over replications (i.e., the standard deviation of the equated scores over hypothetical replications) is sometimes referred to as the standard error of equating. SEE is an example of a random error and it gives the degree of precision with which scores on one test have been transformed to scores on another test (Kolen & Brennan, 2004). To develop its formula with Form X and Form Y as tests to be equated, $\hat{e}q_Y(x_i)$ is defined as an estimate of the Form Y equivalent of a Form X score and E[$\hat{e}q_Y(x_i)$] as the expected equivalent over random samples. Equating error at score x_i is given by: $\hat{e}q_Y(x_i)$ - [$\hat{e}q_Y(x_i)$] and its variance over replication is: $\mathrm{var}\left[\hat{e}q_Y(x_i)\right] = E\left\{\hat{e}q_Y(x_i) - E[\hat{e}q_Y(x_i)]\right\}^2$. The SEE is computed by taking the square root of the variance:

$$se[\hat{eq}_Y(x_i)] = \sqrt{\text{var}[\hat{eq}_Y(x_i)]} = \sqrt{E\left\{\hat{eq}_Y(x_i) - E[\hat{eq}_Y(x_i)]\right\}^2}$$

The magnitude of this expression depends on the type of equating function used, on the data collection design employed in the equating process, and on the method used to smooth the distributions of scores (van Davier, et. al., 2004).

In contrast, there is also systematic equating error, which results from violation of the assumptions and conditions of equating (Kolen & Brennan, 2004). Unlike the SEE, systematic errors are difficult to quantify and the only way to minimize them is to meet the assumptions of the equating procedure and the assumptions of the equating design. Kolen and Brennan (2004) admonish us to minimize any kind of error as much as possible when conducting and designing an equating study.

Kolen and Brennan (2004) recommended two main procedures for estimating SEE: the bootstrap, and the analytic (delta) methods. What follows are steps they for computing bootstrap and analytic standard errors using the random groups design.

1. Draw a random bootstrap sample with replacement of size N_X from the sample of N_X examinees.
2. Draw a random bootstrap sample with replacement of size N_Y from the sample of N_Y examinees.
3. Estimate the equipercentile equivalent at x_i using the data from the random bootstrap samples drawn in steps 1 and 2, and refer to this estimate as $\hat{e}_{Yr}(x_i)$.
4. Repeat steps 1 through 3 R times, obtaining bootstrap estimates $\hat{e}_{Y1}(x_i)$, $\hat{e}_{Y2}(x_i)$, ..., $\hat{e}_{YR}(x_i)$
5. The standard error is estimated by:

$$se_{boot}\left[\hat{e}_Y(x_i)\right] = \sqrt{\frac{\sum_r \left[\hat{e}_{Yr}(x_i) - \hat{e}_{Y\cdot}(x_i)\right]^2}{R-1}}$$

where $\hat{e}_{Y\cdot}(x_i) = \frac{\sum_r \hat{e}_{Yr}(x_i)}{R}$.

The analytic procedure for computing standard errors of equating presented in this section is referred to as the delta method (Kolen & Brennan, 2004). If $eq_Y(x_i;\Theta_1,\Theta_2,....\Theta_t)$ is defined as the equating function of test score x_i and Θ_1, Θ_2, and Θ_t are moments (if equating is by linear method) or cumulative probabilities (if equating is by equipercentile method), the expression for the sampling variance by the delta method is given as: $\text{var}[\hat{eq}_Y(x_i)] \cong \sum_j eq'^{\,2}_{Yj}\,\text{var}(\hat{\Theta}_j) + \sum_{j\neq k}\sum eq'_{Yj}eq'_{Yk}\,\text{cov}(\hat{\Theta}_j,\Theta'_k)$

where $\hat{\Theta}_j$ is an estimate of Θ_j and eq'_{Yj} is the partial derivative of eq_{Yj} with respect to Θ_j and evaluated at x_i, Θ_1, Θ_2,.... Θ_t. The standard error is computed by taking the square root of this variance expression. Kolen and Brennan (2004) proposed the following steps to apply the delta method:

1. Specify the error variance and covariances for each $\hat{\Theta}_j$.
2. Obtain the partial derivative of the equating equation with respect to each $\hat{\Theta}_j$.
3. Substitute the variances and partial derivatives into the expression just presented.

2.6.2 Consequences of Equating

Skaggs (1990) observed that changing the criteria for evaluating equating results changes the conclusion one makes about the adequacy of equating. This lack of consensus needs to be dealt with by finding a suitable criterion for evaluating the results of equating. Assessing the impact a particular equating process has on the sample of

students in terms of the important decisions made based on adjusted scores may be one of the appropriate ways of bringing this consensus. However, this criterion has not been extensively studied. Therefore, more studies investigating the effects of equating on classification decisions are warranted.

Unlike statistics, consequences are not influenced by sample sizes. For example, the standard error of equating (SEE) becomes inconsequential for large samples (Kolen & Brennan, 2004). Although small SEE means better precision, examining the effects of equating on the decisions made based on scores may be a straight forward and easy way of assessing whether equating results are adequate.

2.6.3 The Root Mean Square Difference

Many studies (Eignor et. al., 1990; Gafni & Melamed, 1990; Harris & Kolen, 1990; & Kolen & Harris, 1990) have evaluated the results of equating by applying a series of equating methods to a particular situation to determine whether the methods appear to be providing similar or dissimilar results. Oftentimes, the standardized root mean square difference or its variation is used to indicate the magnitude of the difference between methods. The RMSD discussed here was proposed by Dorans and Holland (2000) and it is a type of an "effect size measure" used after the fact, for assessing the degree to which two equating functions are similar. This index is given by:

$$\text{RMSD} = \frac{\sqrt{\sum_j w_j [ep_j(y) - ep(y)]^2}}{\sigma_{Xp}}$$ where w_j is the relative proportion of examinees in the small sample to a large sample, P is the large sample, p_j is the small sample, $ep_j(y)$ is the equating function for Y to X on p_j; and $ep(y)$ is the equating function of Y to X on P. The denominator σ_{XP} is the standard deviation of the scores of

the large sample. By dividing by this quantity, the measure of RMSD ean be eonsidered as considered an effect size measure, whieh makes this statistic more easily interpretable. This index has been used in this study to compare the random groups design and the external anchor test design.

2.6.4. The Reduction in Uncertainty Index

Dorans (2004) also suggested the use of the Reduction in Uncertainty (RiU) index to help decide whether to ehoose concordance or prediction if the tests to be linked measure different or similar constructs. However, this index is used in this study to help determine the usefulness of the external anchor test in equating. The index is defined as: (RiU) = 1 - CoA = 1 - $\sqrt{1-r^2}$ where CoA, a coefficient of alienation, is a measure of uncertainty about a test that remains after including information from the other test and r is the correlation between the tests. In general, the index is written as:

$$(RiU) = \frac{\sigma_{XP} - \sigma_{XP}\sqrt{1-r^2}}{\sigma_{XP}}$$

where σ_{XP} is the standard deviation of X scores in population, P, of examinees. This index was suggested under the assumption that "the distributions of the scores have either been matched or they are similar enough in shape that a linear relationship is adequate for prediction purposes" (p. 233). Based on this formulation, Dorans (2004) showed that a correlation of 0.866 is required to reduce the uncertainty by at least 50% and that if a predictor cannot reduee the uncertainty by this much, "it is unlikely that it can serve as a valid surrogate, via concordance or equating, for the score being predicted" (p. 231). In this study, this criterion was used to make judgment regarding the usefulness of equating test forms.

2.7 Summary of the Review

In summary, this chapter has reviewed studies, papers and books relating to necessity and appropriateness of equating, steps in the equating process, equating designs, equating methods, equating using anchor tests, and ways of evaluating the results of equating. Mainly, this review serves to provide this study a theoretical grounding and to inform the methods for data collection and data analysis. It is important, however, to highlight important aspects of the review and show how the present study intends to contribute to this body of knowledge.

The review has highlighted the importance of beginning the process of equating by evaluating the test forms and deciding, based on the information collected, whether to go ahead with equating. This is because not all tests can be equated and equating tests whose distributions are very different or very similar can add in more error than what the process itself intends to remove. The evaluation framework proposed by Kolen (2004) or that by Dorans (2004) discussed in this chapter could be used to determine whether equating is appropriate. Tests that measure different constructs in different populations, tests whose scores are used to draw different inferences, and tests that have different measurement conditions should not be equated because their score distributions are often very different. The identity equating and the more objective log-linear methods can be used to evaluate whether tests have very similar score distributions to warrant equating unnecessary. Since it is known that test forms are rarely strictly parallel to produce very similar distributions, most testing agencies do not bother checking whether equating is necessary. Tests with very similar score distributions should not be equated. However, testing boards that intend to start equating test forms should embrace this idea.

The steps outlined by Kolen and Brennan (2004) can guide testing agencies during the equating process because they include most of the steps proposed by other researchers such as Harris and Crouse (1993). An examination board can choose an appropriate equating design and an appropriate equating method from the ones discussed in this chapter. Most testing agencies in the industry, however, use the NEAT design because their test forms are usually taken by different populations of students. In this study, two designs, the random groups and a modified NEAT design, the external anchor test design, were used to collect data from two groups of students. Whatever equating design is used, evaluating the adequacy of equating is very important because it tells us whether the process has been successfully executed. Procedures for evaluating equating have been reviewed ranging from traditional standard errors to classification of examinees.

Studies about consequences of not equating educational tests were not found. Instead many equating books and papers assume that people are aware of what happens when test forms are not equated. Oftentimes, differences in test difficulty and invariance of examination standards are cited as some of the reasons for requiring test forms to be equated. But they do not explicitly show, using empirical examples, that examination standards vary when test forms across years are not equated. Similarly studies that looked at the effect that equating has on the classification of students have not been spotted. The uniqueness of this study rests on the fact that it launched an empirical investigation into these issues. Examining the effects of not equating high stakes educational tests in terms of important classification decisions made based on scores can help to make a case for equating.

The review has shown that common items (internal or external), different tests, and other school variables could serve as anchor. Whenever, common items are used as anchor, they are often drawn from the reference form. There has been very little research that has investigated the possibility of equating test forms using anchor items that are not coming from the reference form. The present study was intended to make a contribution towards this end. It is designed to investigate the possibility of using external anchor items that are constructed by teachers every time test forms are to be equated. Knowledge in this area is important for examinations boards that disclose all items in the reference test form after administration for instructional and security reasons.

The review has also shown that anchor tests and operational test forms are usually administered concurrently. Both the common items and different tests are usually given to examinees together with a very short or zero time interval between the administrations. However, it would be interesting to investigate whether equating would be adequate when anchor tests are administered separately from the operational tests. Knowledge in this area would be important in situations where there are concerns of over burdening examinees during the operational tests. Such concerns usually exist in examinations boards like MANEB that handles high stakes examination. This study intends to make a contribution towards this end.

Whenever different tests and school variables serve as anchor, they tend to be dissimilar in content from the target tests. From the unitary validity perspective, the target and anchor tests in these instances measure different constructs. It is sounds unreasonable and theoretically unappealing, to this researcher, to have examinees' differences on one construct serve as a basis for adjusting possible group differences on another construct.

Therefore, developing a short external anchor test that is similar in content to the operational tests may be a good alternative. This study intends to build on this body of knowledge and help to inform dialogue regarding equating using separately administered anchor test.

CHAPTER 3

METHOD

In this chapter, methodology for the study is described. It includes the descriptions of the participants, instruments, the proposed data collection design, and procedures for data analysis. The first part of the data analysis section explains the preliminary analyses that were carried out to understand the data and analyses that were conducted for purposes of choosing the appropriate smoothing method.

3.1 Participants

The study took place in Malawi and the participants included students, teachers, and MANEB. This section describes the characteristics of these participants that took part in the study.

3.1.1 Schools and Students

The study used scores for a sample of 1,017 eighth grade students who were enrolled in 12 Primary Schools in Zomba District for the 2005 academic year. A stratified random sampling procedure was used to draw this sample of schools from the population of 89 primary schools in Zomba district and all the eighth graders in the sampled schools were involved in the study. Of these participating students, 53% were girls and 47% were boys. Of the 12 schools, 6 were in urban areas whereas the other 6 were in rural areas; 2 were girls' schools, 2 were boys' schools, 8 of them were co-education schools enrolling both sexes, 4 of the schools in the sample were mission schools run by religious leaders whereas as 8 were government schools run by the Ministry of Education. There were five pilot schools (3 urban and 2 rural) randomly drawn from Zomba with an estimated sample of 369 students. The pilot schools were not part of the operational sample.

3.1.2 Teachers

The study also involved a group of 14 mathematics teachers who were teaching the grade eight students in the sampled schools to administer and score the tests. These were qualified teachers with MSCE certificates and professional teaching (PT2) certificates from the Malawi National Examinations Board (MANEB). They had a minimum of 8 years and a maximum of 22 years of teaching experience. Each one of them was a trained scorer by MANEB who had been scoring national examinations for not less than 4 consecutive years. All of them were, at the time of the study, still actively involved by the board in scoring mathematics examination scripts. Of these teachers, 5 were female and 10 were male; 2 were senior examiners for national examinations whereas the other 12 were just experienced scorers. The senior examiners are usually responsible for training scorers, supervising marking of exams and writing examiner's report to the chief examiner regarding the marking exercise.

The study also involved 12 head-teachers from the schools that were in the sample to motivate students and to supervise the administration of the tests. These were highly experienced individuals with more than 10 years experience as administrators of primary schools. At the time of this study, they already had the responsibility to supervise the administration of the mock exams that were going on in their respective schools. They were very instrumental in allowing the tests to be administered as part of the mock examinations.

Another group of 5 teachers from schools other than those in the sample (pilot schools) constructed a set of anchor items. These too were qualified teachers holding MSCE and PT2 certificates with more than 10 years of teaching experience. They were

highly trained scorers with more than 8 years scoring experience and they were, at the time of this study, still involved in marking MANEB exams. Of the five teachers 4 were males and 1 was female; 2 were deputy head teachers in their respective schools whereas the other 3 were just math teachers.

3.1.3 MANEB Officials

Finally, a group of six MANEB officials participated in the "Awards Meeting" to set cut scores on the tests. All six were highly experienced individuals heading important departments and sections in the Board. One person in the group had a doctorate degree in educational measurement whereas the rest had a masters degree. These officials were a subset of the larger committee of 10 – 12 people that sets cut scores on operational PSLCE tests. Of these, 3 were directors of important departments in the Board chosen for their relevance to the study, 2 were section heads also chosen for their relevance to the study and 1 was subject officer for mathematics.

3.2 Instruments

3.2.1 Test Forms

The study used the 2003, 2004 and 2005 Primary School Leaving Certificate Examination (PSLCE) mathematics tests constructed by the Malawi Examination Board (MANEB) to collect students' performance achievement. The 2003 test comprised of 30 multiple choice items and 10 constructed response items with a total of two hours of testing time whereas the 2004 test comprised of 30 multiple choice items and 7 constructed response items with a total of two hours of testing time. The 2005 test also had 30 multiple choice items and 8 constructed response items. The maximum score on each of these test forms was 100 score points.

The 2003, 2004, and 2005 tests are alternate forms designed to measure the same construct - students' mathematics proficiency at the end of the primary school cycle. Parallelism among these forms is sought by ensuring that each form assessed the same content areas. These areas are spelled out in the Malawi Primary School Mathematics Teaching Syllabus (Malawi Ministry of Education and Culture, 1991) as number and numeration; money; geometric shapes; measurement; graphs; rate, ratio, and proportion; postal services; bank services; and simple accounts. The test forms are administered under similar conditions and their resulting scores are used by the Ministry to support certification and selection decisions. Therefore, the test forms can be equated because they satisfy the evaluation criteria proposed by Kolen (2004). The choice of these test forms was also deliberate in that they are more recent tests.

3.2.2 External Anchor Test

The study also used scores on an anchor test. Given that all the items on 2003 and 2004 tests were disclosed, they could not be used as common items to form an equating anchor test. Instead, a new set of items were created and used as an external anchor test. This set of items was constructed by a group of five mathematics primary school teachers who were given a brief orientation in item writing. They were then asked to write items that were similar in content and format as those on the 2003 and 2004 tests making their short test a mini version of the two tests. From these items, 12 items (9 multiple choice and 3 constructed response items) were selected to constitute the anchor test and as such the length of the anchor test was 20% of the total test. The test was pilot tested and using the scores on the pilot exercise, the characteristics of anchor items were gathered. This information was instrumental in revising the set of external anchor items making them as

highly comparable to the operational tests as possible. During the assembling of the anchor test, recommendations discussed in the literature review chapter of this dissertation such as difficulty level of items, length, similarity in content and format were considered to promote its quality.

3.2.3 Survey Items

It was important in this study that students remain as motivated as they would have been during the operational MANEB tests otherwise the results would be greatly confounded. Therefore, the researcher created 3-survey questions (Likert type) asking students to indicate, on a 5-point scale, how they considered the importance of the tests, how prepared they were for the test, and how hard did they try to respond to the items on the tests. The list of these questions is given in Table 3.1

Table 3.1: Survey Questions

Question	Scale
1. How important is this mock exam to your preparation for the MANEB tests?	From 1=Not at all Important To 5=Very Important
2. How prepared were you to write this exam?	From 1=Not at all Prepared To 5=Very Much Prepared
3. How hard did you try to answer the questions on this test?	From 1=Never Tried Hard To 5 = A Great Deal Hard

This survey was intended to give the researcher a signal regarding the degree of motivation, which the students had during the re-administration of the tests. These items too were pre-tested in the 5 primary schools (pilot schools) described in the preceding section prior to data collection exercise. Information gathered from this pilot exercise was used to revise the items.

3.3 Data Collection

3.3.1 Data Collection Design

Two equating designs, the random groups design and an external anchor test design were used in this study to collect data for equating test forms. Table 3.2 summarizes the data collection design in the study.

Table 3.2: The Data Collection Design

	1st Administration (July 25, 2005)	2nd Administration (July 27, 2005)	3rd Administration (September 1, 2005)
Group A	2004 Test	Anchor Test	2005 Test (MANEB)
Group B	2003 Test	Anchor Test	2005 Test (MANEB)

In the random groups design (see Table 3.2), two independent, random samples of examinees (Group A and Group B) were obtained through random assignment of the students from the 12 schools. Group A had 506 students whereas Group B had 511 students and the 2004 test was administered to group A while group B took the 2003 test. This first administration occurred on the 25th of July 2005. To ensure equivalency of the groups, the spiraling process was employed during the administration of the tests. In this process, the first candidate in each class received the 2004 test booklet, the second candidate received the 2003 test booklet, the third examinee the 2004 booklet, and so on. This kind of spiraling is known to produce comparable, randomly equivalent groups (Kolen & Brennan, 2004).

In an external anchor test design, the whole sample of 1,017 students took the anchor test during the second administration which occurred on the 27th July 2005 and later they took the 2005 operational test during the third administration on the 1st of September, 2005. Note that the time interval between the first and the second

administration was 2 days whereas the time interval between the second and the third administrations was 5 weeks. It was hoped that 5 weeks was short enough to minimize the learning effects that would otherwise have taken place between the time students took 2004, 2003 and anchor tests and the time they wrote the operational test. An attempt was also made to establish whether or not students have seen or taken 2004 and 2003 test forms prior to the mock time through interviews with teachers and students.

3.3.2 Administration of the Tests

As mentioned in the preceding sections, students' motivation was crucial in this study and as such, the 2004, 2003 and the anchor tests were administered as mock (practice) examinations to cultivate students' interest and attain the motivation that would be comparable to what would be there during MANEB examinations. In case of Zomba Schools, mock tests are customary and usually students write two different sets of mock exams – one for the zone and the other for the district. The 2004, 2003 and the anchor tests in this study were administered as one of the district mock exams. Prior to the exercise, the head-teachers worked to motivate students in their schools to take the mock test seriously knowing that it would help them prepare for the operational exams. The 2005 test was administered by MANEB and this researcher had no control over the exercise. Nevertheless, the administrative conditions for all the tests were similar.

Logistically, the 2004, 2003, and the anchor tests were all administered by the selected 12 mathematics teachers. To minimize examination malpractices arising from teachers' laxity and from teachers helping their students, each teacher was assigned to a different school to invigilate the exams. Before the exercise, all participating teachers were briefed on the administrative conditions for the tests. Head-teachers supervised the

administration of the tests in their respective schools to ensure that rules were being followed. After invigilation, teachers were required to report any problem encountered during the administration exercise to be included in the examiners' report.

3.3.3 Scoring

The 2004, 2003 and the anchor tests scripts were scored by the 12 mathematics teachers who also took part in the administration process. The whole marking exercise was supervised by 2 MANEB senior examiners who also wrote a report that later formed part of the information used for setting cut scores. As described earlier, all teachers and senior examiners involved were experienced and highly trained scorers.

Marking took place at one place (The Teacher Development Center). All the teachers were brought together at this marking center for a day to mark the scripts. Organizing it this way seemed important for a number of reasons: (1) it accorded teachers an opportunity to standardize the scoring rubrics and ask one another any question they might have prior to the marking time; (2) it helped to minimize the tendency by teachers to inflate scores for their students, and (3) it helped to facilitate spot checking of the marked scripts by the supervisors to ensure that they were being reliably scored. Teachers were not allowed to score papers from their own schools. All attempts were made to make the marking process as close to the MANEB process as possible.

MANEB's nominal registers also were screened to identify repeaters after failing to identify such individuals through interviews with teachers. Majority of those identified were external examinees (i.e., examinees who were not attending classes in the regular schools, but were only there to take exams). Teachers reported that there were no repeaters among their students and students themselves self-reported that they were

attending grade eight for the first time. Later the researcher compiled the composite as well as the item-level scores for each examinee into an SPSS file database. A few months later, scores on 2005 math test for the participating students were obtained from MANEB. Scores for the repeaters on all test forms were not filed.

3.3.4 Setting Cut Scores

The Cut scores on the 2004, 2003 and the 2005 test forms were set by the MANEB officials during the "Awards Meeting." Six people (60% of operational committee) participated in the process. During the meeting different kinds of information were used to inform the process. Among others, the committee considered a report from the senior examiners describing what happened during the administration and scoring exercises. They also looked at the item statistics and the distribution of scores to take note of their levels of difficulty, and test items themselves to see whether they contained typos or spelling errors. Using this multifaceted information, a heated debate then followed with suggestions from members regarding what would be the appropriate grade boundaries (cut scores) for A, B, C, D, and F performance categories. The final cuts were arrived at through consensus among members.

3.4 Preliminary Analyses

This section describes the preliminary analyses that were carried out in this study. Some of these analyses (item and reliability analyses) were conducted to understand the technical quality of the test forms as well as the quality of the external anchor test while others were carried out to test important assumptions such as the motivation of students and group equivalency. The information from the preliminary analyses aids in interpretation of the results of the study.

3.4.1 Item and Reliability Analyses

It is important to understand the data before any analysis is done. Statistical properties of the item scores such as the difficulty and the discrimination indices were examined through item analysis. The reliability of 2004 and 2003 test forms were also computed to determine how reliable were the test scores to be used in equating processes. The multiple choice section on each test form is weighted differently than constructed response section when computing the composite score for examinees. As such, the alpha values for multiple choice and constructed response sections on each test form were estimated separately. The Nunnally (1967) formula was used in the study to estimate the reliability of the weighted sum. The formula is:

$$r_{yy} = 1 - \frac{\sum b^2_i - \sum b^2_i r_{ii}}{\sigma^2_y}$$

in which b_i^2 is the weight of the section i, r_{ii} is the reliability of the section i, and σ_y^2 is the variance of the linear combination of scores for the sections.

3.4.2 Choosing a Smoothing Procedure

The study used equipercentile equating procedures to equate scores on the test forms. Whenever, this definition of equating is used, the equipercentile relationships tend to be irregular because of the random error in estimating the equivalents (Kolen & Brennan, 2004). It is customary, therefore, to employ smoothing methods to obtain regular distributions and less random error. In this study, two presmoothing methods (log-linear and beta4 methods) and one postsmoothing method (Cubic Spline method) were explored and compared in the interest of choosing an appropriate model to use in whenever equating is done.

The log-linear method was used to pre-smooth the raw score distribution by fitting the following model to both Form X and Form Y data.

$$\log[N_X f(x)] = w_0 + w_1 x + w_2 x^2 + \ldots\ldots + w_C x^C$$

In this model, $\log[N_X f(x)]$ is log of the density expressed as a lower order polynomial of the degree C. The choice of C was aided by inspection of the moments, the likelihood ratio chi-square goodness-of-fit statistics, and by inspection of graphs.

The beta4 method (Lord, 1965 in Kolen & Brennan, 2004), a strong true score procedure, was used to presmooth the observed score distributions. The method assumes a true score distribution, $\psi(\tau)$, and a conditional observed score distribution given true score, $f(x \mid \tau)$, such that the observed score function is expressed as follows:

$$f(x) = \int_0^1 f(x \mid \tau)\psi(\tau)d\tau$$

In the interest of fitting a wide variety of shapes, Lord proposed that the true score distribution should be a four-parameter beta with two parameters whereas the conditional distribution should be a compound binomial. The beta4 distribution, $f(x)$, is estimated using information regarding the number of items, the first four moments of the sample distribution, and Lord's k parameter. In this study, the Lord's k was set at zero and only the first three moments were fit to the data because fitting all the first four moments resulted into upper limits for proportion-correct true scores that were above one.

The cubic spline method, also described by Kolen and Brennan (2004), was used to directly smooth the equipercentile equivalent, $\hat{e}_Y(x)$. In this postsmoothing procedure, the following continuous spline function is fit to each score point:

$$\hat{d}_Y(x) = v_{0i} + v_{1i}(x - x_i) + v_{2i}(x - x_i)^2 + v_{3i}(x - x_i)^3, x_i \le x < x_i + 1.$$

where v_{0i}, v_{1i}, v_{2i}, and v_{3i} are weights that take on different values at each score point resulting into each integer score having a different cubic equation. The function is fit over a range of scores x_{low} to x_{high}, $0 \le x_{low} \le x \le x_{high} \le K_X$, where x_{low} is the lower score point in the range and x_{high} is the upper score point in the range. For those integer scores where the frequency is zero, a linear interpolation procedure is used to obtain equipercentile equivalents outside the range of the spline function. The summation of the spline functions over score points is minimized subject to satisfying the following constraints:

$$\frac{\sum_{i=low}^{high}\left[\frac{\hat{d}_Y(x_i)-\hat{e}_Y(x_i)}{\hat{se}\left[\hat{e}_Y(x_i)\right]}\right]}{x_{high}-x_{low}+1} \le S$$

The term $\hat{se}[\hat{e}_Y(x_i)]$ is the estimated standard error of equipercentile equating. The S parameter controls the degree of smoothing. In this study the choice of S was aided by inspection of moments and graphs.

Finally, the model for smoothing a particular equating relationship was selected by comparing the models of choice among the log-linear, beta4, and cubic spline. The criterion was to select a method that results into a smoothed equipercentile distribution that appeared smooth enough without departing too much from the observed unsmoothed relationship.

3.4.3 Students' Motivation

The responses to the survey questions were analyzed by computing the percent of examinees who indicated that they saw the tests as important, and that they tried their best to respond to the items on the tests. The regression analysis was also carried out to

explore the extent to which variables such as importance, preparedness and trying hard predict achievement on the tests. These analyses gave us a rough picture regarding the degree of motivation for the group of students participating in the study.

3.4.4 Establishing Group Equivalency

The groups that took the tests were presumed to be equivalent because a spiraling procedure was used during the administrations of the tests. However, it was necessary to check this presumed equivalency of the groups after data collection. In this analysis, the mean scores on the external anchor test for the two random samples (defined by test form) were compared using the independent t-test. Another common measure for the groups was the 2005 test form and the difference in means scores on this form too was tested through the independent t-test.

3.5 Data Analysis

The study used linear and equipercentile equating methods to equate scores from the random groups design and from the external anchor design. These procedures are described in detail by Angoff (1971), Braun and Holland (1982), Kolen and Brennan (1995, 2004), and Livingston (2004). A brief discussion of the equipercentile methods was also given in chapter 2 of this dissertation.

The following equatings were carried out at various stages of the analysis: (a) random groups equipercentile equating of the 2004 to the 2003 test, (b) external anchor test equipercentile equating of the 2004 to the 2003 test, (c) random groups equipercentile equating of the 2005 to the 2004 test, (d) random groups equipercentile equating of the 2005 to the 2003 test, (e) external anchor test equipercentile equating of the 2005 to the 2004 test, (f) external anchor test equipercentile equating of the 2005 to the 2003 test, (g)

Tucker Linear equating of the 2005 to the 2004 test, and (h) Tucker Linear equating of the 2005 to the 2003 test. Some of these analyses were carried out for purposes of comparing the results. There were other analyses too designed to investigate each of the research purposes appearing in chapter 1.

3.5.1 Is it Necessary to Equate these Tests?

The question was answered by comparing the difficulty levels and score distributions on 2004 and 2003 test forms. Plots and statistical tests were instrumental in these analyses.

3.5.1.1 Comparing the Difficulty of the Tests

The descriptive statistics for the 2004 and 2003 test forms were obtained to examine the difficulty of the two tests. By using the randomly equivalent groups design, any difference between the group-level performances on the two tests could be taken as a direct indication of the difference in difficulty. Therefore, to establish which test form was relatively more difficult than the other, differences in means on the 2003 and 2004 tests for the two groups were tested using the independent t-test.

3.5.1.2 Comparing Score Distributions of the Tests

A number of procedures were used to make the comparison of the score distributions. Fist, the score distributions of students on the 2003 test were compared to the students' score distribution on the 2004 test by plotting them on the same axes to graphically establish the extent to which the two distributions were similar. With such a plot, it was easy to notice how close the distributions are located on the score scale. Hanson's (1992, cited in Kolen & Brennan, 2004) log linear method was used to investigate whether the distributions of these two tests were similar enough to warrant

equating unnecessary. In this analysis, the chi-square test was used to test the null hypothesis that the distribution of raw scores on the 2003 and 2004 tests were similar. This procedure was used to bring in objectivity in the process of comparing score distributions.

Secondly, the equipercentile equating line for 2004 to the 2003 test was compared to the identity equating line plotted on the same axis. To accomplish this, scores on the 2004 test were equated to scores on the 2003 forms using the post-smoothed equipercentile equating method (Angoff, 1971; Kolen & Brennan, 1995 & 2004). A computer program called RAGE-RGEQUATE (Cui & Kolen, 2005) was used to estimate the smoothed equipercentile relationships. During equating, the 2003 test was treated as the reference (old) form whereas the 2004 test was treated as the new form. The Dorans and Lawrence (1990) criterion was used in making the comparisons of the equating lines. Using this criterion, the identity line that falls within ± 2 standard errors after equating indicates that equating is unnecessary. This evidence was necessary to support earlier evidences for or against equating.

3.5.2 Invariance of Examination Standards

In this study, the invariance of the examination standards on the tests was investigated by comparing the cut scores on the 2004 and the 2003 test forms before equating. Scores on the 2003 and those on the 2004 test forms were also linearly transformed to obtain z-scores by subtracting their respective group means from each score and divide the difference by the corresponding standard deviations of the scores. The cut scores were compared by looking at how far they are from their respective group means in standard deviation units.

3.5.3 Effect of Equating on Examinees' Classification

The effect of equating on examinees' classification was investigated through decision consistency analyses and by comparing the classification of students into different grade boundaries on the 2004 and the 2003 test forms before and after equating. In the classification exercise, students were classified into A, B, C, D, and F categories based on the unadjusted 2004 and 2003 scores. Later the same students were classified into the categories using the adjusted 2004 scores. The percentages of students in the categories before and after equating were compared. In the decision consistency analyses, the pass/fail decisions based on 2004 and 2005 test scores before equating were compared to similar decisions after equating the two test forms. Similarly, decisions based on the 2003 and 2005 scores were compared before and after equating. The decision consistency (DC) indices were computed in all cases to index the consistency in decisions. However, this index does not take into account consistencies in decisions due to chance and as such Cohen's kappa ($_K$) was computed along side the decision consistency index to facilitate interpretation of the results.

3.5.4 Equating Using External Anchor Test

One of the goals in this study was to investigate the effectiveness of using an external set of items (anchor) that is administered during mock (practice) period to equate the operational tests. To accomplish this goal, it is important to establish that students' abilities do not change significantly from the time they write mock exams (which included the external anchor test) to the time they write the operational MANEB tests. Furthermore, it should be established that such an anchor test can provide adequate information to be useful in equating. The procedures described in this section were

employed to rule out learning effects and to determine whether the external anchor test that is given separately from the target tests had the potential to provide adequate information to be useful in equating. Procedures for equating test forms through the anchor are also described.

3.5.4.1 Were there Significant Learning Effects?

It was not possible to rule out learning effects with the available information and from the classical measurement perspective. However, the anchor test was administered 5 weeks before the operational 2005 tests. This time interval was considered short enough to minimize the learning effects between the two administrations. Since the anchor test was administered almost at the same time (with an interval of 2 days) as 2004 and 2003 forms, performance of candidates on 2005 were compared to their performance on 2004 and 2003 tests to establish if students changed significantly from the time they wrote mock exams to the time they wrote the operational MANEB tests. The compared was made by testing the difference between the group means through the paired samples t-test.

3.5.4.2 How Useful is the Anchor Test?

Note that the group of students who took the 2004 test, the 2003 test, and an anchor test were the same students who took the 2005 test. Therefore, using the correlation between the scores on anchor test and scores on the 2005 form, the reduction in uncertainty index was computed to determine if information from scores on anchor items that were administered apart from the operational test has the potential to reduce the uncertainty of knowing the students' performance on the 2005 test. The index was also computed for scores on the 2003 and the 2004 tests versus scores on the anchor test

and for 2005 group A and 2005 group B scores for purposes of comparison. The criterion proposed by Dorans (2004) that RIU should be greater than equal to 0.87 in order for the anchor test to be deemed useful was used to aid interpretation of the results.

Again, the following equatings were carried out: (a) the random groups equipercentile equating of 2004 test to 2003 form, (b) the random groups equipercentile equating of 2005 test to 2004, (C) the random groups equipercentile equating of 2005 test to 2003, (d) the external anchor test equipercentile equating of 2004 test to 2003 form, (e) the external anchor test equipercentile equating of 2005 test to 2004, and (f) the external anchor test equipercentile equating of 2005 test to 2003. The results from the random groups equating were compared to their corresponding external anchor equipercentile equating results. The root mean square difference proposed by Doran (2004) was used to index the difference between the equating relationships. Comparing the designs in this way would help to determine to extend to which results from the external anchor test design deviate from the results from the random groups design. Such information would in turn reveal the potential that the external anchor test has to provide useful information that can be used to equate the 2005 test to the 2004 test.

3.5.4.3 Equating via the Anchor Test

The scores on 2005 test were equated to scores on 2004 and 2003 scores through the external anchor design. The postsmoothed frequency estimation method (an equipercentile method of equating) and the Tucker linear equating method were employed in this process to equate the scores on the test forms. The equating was carried out using a computer program called CIPE (Kolen, 2003). The mean squared equating errors for the equatings were computed to establish the adequacy of equating.

Finally, the classification of students into pass/fail categories (pass rates) were compared before and after equating the 2005 test to 2004 and 2003 forms. The magnitude of the difference in pass rates was used to facilitate the comparison. This analysis was carried out to establish whether equating through the anchor test still improves the decisions made based on test scores. It was one way of establishing the adequacy of equating.

3.6 Summary of the Method

In summary, this study engaged 1,017 grade eight students from 12 primary schools around Zomba district in Malawi. The tests that were equated were the 2003, 2004 and the 2005 PSLCE math exams prepared by MANEB. Random groups and external anchor test designs were used to collect data and both the postsmoothed equipercentile and the Tucker linear equating method were used to equate test forms. The equated results were used to create conversion tables. However, before equating began, it was important to investigate the level of difficulty for the test forms, find out whether it was necessary to equate them, and whether the cut scores remain invariant. These investigations were conducted in line with the purposes of the study.

The difficulty of the tests was examined by comparing group means through the independent t-tests and the necessity to equate tests was determined by comparing the score distributions using log linear and identity equating methods. The invariance of the examination standards was determined by comparing the cut scores set by the "Awards Meeting" on each test form. The effect of equating on examinee classification was investigated by comparing the classification of students into different grade categories after and before equating and through decision consistency analyses.

The study used the Reduction in Uncertainty Index (RIU) and comparison of equating results across methods to establish if, in fact, an external anchor test would be useful in equating. The standardized root means square difference (RMSD) was used to index this difference between the equating methods. Finally, standard errors of equating were very instrumental in all the equating processes to evaluate the adequacy of the equating results and to facilitate comparison of equating relationships.

CHAPTER 4

RESULTS

This chapter presents the results of the study. The first part of the chapter presents the results from the preliminary item and reliability analyses. These include item discrimination, item difficulty, and reliability indices for items and test forms as well as for the external anchor test. Summary statistics for students' ratings on the survey items, results of the significance testing of the means on the anchor and the 2005 tests are also given as part of the preliminary analysis results. Then model selection parameters and graphs are shown, which were used to select the smoothing procedure.

After the preliminary results, the chapter presents results from the main data analyses addressing the purposes of the study. These include group statistics on the 2004 and 2003 test forms for testing the difference in performance of students on the tests to investigate differences in test difficulty of the test forms; and the plots of score distributions on 2004 and 2003 test for investigating whether their score distributions were different. Next, cut scores on the test forms from the Awards Meeting are presented and comparisons are made, differences are qualified and the conversion tables from the equating of 2004 to 2003 test forms is then given. The tables indicating percentages of students classified by equated scores and scores that are not equated follow and these are used to investigate the effect of not equating scores on examinees classification. This category of results also includes pass rates that provide further information regarding the consequences of not equating educational tests. The chapter then proceeds by presenting results on equating test forms using an external anchor test. They include paired-sample t-tests, the reduction in uncertainty indices, plots showing the equating function from the

random groups design and external anchor test designs plotted on the same axes to facilitate comparison. These results are intended to help in the investigation of usefulness of the anchor test in the equating process and in ruling out learning effects. The conversion tables from equating of 2005 to 2004 and 2005 to 2003 test forms resulting from both equipercentile and linear equating processes are presented last. The chapter ends with a summary of the findings.

4.1 Results from Preliminary Analyses

4.1.1 Item and Reliability Analyses

In preliminary analyses, the tests were statistically evaluated to determine their item and test quality. The corrected item-total correlation, mean (p-values), and alpha values from the item and reliability analyses of the 2004 and 2003 test forms are presented in Table 4.1 whereas the item-total correlation, mean, and alpha values for the anchor test are given in Table 4.2.

4.1.1.1 Item Discrimination

The corrected item-total (point-biserial) correlations for 2004 test items ranged from 0.00 to 0.34 with an average of 0.15 ($S_X = 0.08$) and for 2003 items they ranged from 0.01 to 0.39 with an average of 0.18 ($S_Y = 0.13$). In general, items in both forms had modest point-biserials indicating that they would not highly discriminate students who possess the characteristics of interest from those who do not have such characteristics. Particularly, items 1, 13, 18, 21, 22, 24, 25, and 27 on 2004 and items 2, 8, 10, 12, 13, 14, 20, 23, 26, 27, and 29 on 2003 had zero or near zero point-biserials, which implies that they do not contribute much to the measurement done by other items. Note also that all items had positive discrimination values on 2004 whereas on 2003 items 2, 8, 23, and 26

had negative discrimination indices. These later items seem to be working against the discrimination done by other items. However, examination of the items indicates that they were not miskeyed and deleting them does not seem to improve the reliability of the test.

The point-biserial correlations for the anchor test items ranged from 0.08 to 0.26 with an average of 0.17 ($S_A = 0.07$). The values are very close to those of 2004 and 2003 test forms. These observed moderate point biserial correlations imply that the external anchor items too were not highly discriminating between students who possessed the characteristics of interest from those who did not have such characteristics. This finding may not be very surprising because the anchor test was constructed in such a way that it resembles the test forms. Two items (4 & 5) had near zero point biserials indicating their minimal contribution to the measurement done by other items. These items, however, could not be taken out because they were assessing important content areas.

4.1.1.2 Item Difficulty

The p-values for multiple choice (MC) items on 2004 test form ranged from 0.07 to 0.77 whereas those of constructed response (CR) items ranged from 0.03 to 0.50. On 2003 test form, p-values for MC items ranged from 0.16 to 0.77 whereas those of CR items ranged from 0.11 to 0.48. The average item mean for the 2004 test form was 0.30 ($S_X = 0.19$) whereas the average for the 2003 form was 0.37 ($S_Y = 0.17$). Items 2, 3, 12, 14, 28, 29, 33, and 35 on 2004 were easy items with high p-values ($p > 0.50$) and on 2003 easy items were 1, 9, 17, 22, 24, 25, and 28. Generally, therefore, most items on both forms were difficult with low p-values indicating that most students, including the more capable ones, would not answer them correctly. It was not surprising, therefore, that all

these items were not highly discriminating among students who had mathematics skill from those who did not have mathematics skills. This is an important finding in that when test are more difficult guessing increases and this in turn affects the reliability of the test form.

The MC items on the anchor test had p-values ranging from 0.08 to 0.85 and the two CR items had item means of 0.19 and 0.61. The mean difficulty value for the test was 0.40 ($S_A = 0.25$). For the MC items on the anchor test the p-values were within the same range as those on the 2004 and 2003 test forms with item 4 being the most difficult item. However, the CR items were closer in difficulty to the 2003 test than they were to the 2004 test form with item 11 being the easiest on the test.

4.1.1.3 Reliability

The alpha value for multiple choice items alone on 2004 was 0.53 whereas the CR items on 2004 had an alpha of 0.45. The MC items on 2003 test had alpha of 0.56 whereas the CR items on the 2003 test form had an alpha of 0.40. The anchor test had even lower values of alpha. The MC items on anchor test had an alpha value of 0.37 whereas CR items had an alpha value of 0.36. Given that the MC section of the tests was weighted 0.60 and the CR section was weighted 0.40, the reliability of the total score was estimated using Nunnally's (1967) formula for the reliability of linear composite. For 2004 test the weighted coefficient was 0.63 and that for 2003 was 0.67. The overall reliability coefficient for the external anchor test was 0.49. All these are generally low values and from the internal consistency perspective, they indicate that items comprising the tests were only moderately working together to measure the construct of interest. Of course, the low values of the anchor test should be cautiously interpreted because the test

length was very short and this was one of the sources of its unreliability. For items in each form, the alpha when item deleted were similar, therefore, no single item in both tests highly influenced the reliability estimate for the respective total score.

The relationships between test forms and the anchor test were similarly low ranging from 0.51 (anchor versus 2005 test) to 0.56 (anchor versus 2003 form). These correlations are presented in Table 4.3. The low coefficients signify existence of only moderate relationships between the anchor test and test forms and moderate relationships between the multiple choice items and the constructed response items in the 2004 and 2003 test forms. This finding implies that the external anchor test could explain 26% and 31% amount of variability in 2004 and 2003 test forms respectively.

4.1.2 Choice of the Smoothing Models

4.1.2.1 Choosing a Log-linear Model

The choice of the lower-order polynomial of degree C, for a log-linear model was aided by inspection of graphs presented in Figures 4.1, 4.2, and 4.3 and by checking whether or not moments are preserved as presented in Table 4.4. The goal was to identify the degree of smoothing that does not depart too much from the unsmoothed function. Therefore, different smoothed relationships ranging from C = 1 to C = 10 were compared to the unsmoothed line. The three difference plots presented in this chapter only represent relationships for C = 1, C = 4, and C = 6. The results suggest that the following minimum model with parameter C = 4 fits both 2004 and 2003 score distributions well enough:

$$\log[N_X f(x)] = w_0 + w_1 x + w_2 x^2 + w_3 x^3 + w_4 x^4$$

For instance, inspection of the plots show that the log-linear model with C = 1 results in a well smoothed distribution, but its smoothed function does not approximate the

unsmoothed distribution well enough and it falls outside the plus and minus one standard error band along the entire scale. The polynomial with C = 6 over fits the distributions such that the smoothed line becomes more irregular than it would otherwise be desired. However, with C = 4 the smoothed distribution is not as good as that of C =1 and not as irregular as that of C = 6, but it falls within plus and minus one standard error band for the most part of the scale. It appears, therefore, that for this study, a log-linear model with C = 4 would work better than other models. Inspection of moments in Table 4.4 shows that with C = 4 all the first four moments are preserved both for 2004 and 2003 score distributions.

4.1.2.2 Fitting the Beta4 Compound Binomial Model

When the four beta compound binomial model was fitted to the score distributions, the resulting moments for smoothed distributions presented in Table 4.5 were obtained. The first three moments for both 2004 and 2003 score distributions were preserved. The difference plot in Figure 4.4 shows that the smoothed relationship falls within plus or minus one standard error of equating (± 1SE) band for the most part of the distribution. However, like the log-Linear model, the beta4 method performs poorly on the high end of the distribution.

4.1.2.3 Choosing the Cubic Spline Model

The cubic spline function was used to smooth the equipercentile relationship at various degree of smoothing ranging from S = 0.10 to S = 1.00 and the results were compared through inspection of moments shown in Table 4.5 and through inspection of graphs in Figures 4.5, 4.6, and 4.7. When S = 0.10 only the first two moment (mean and standard deviation) are preserved (to one decimal place). On the other hand, when S =

1.00, all the first four moments (mean, standard deviation, skewness, and kurtosis) are preserved. Intermediate values of S lead to preservation of only the first three moments. However, the difference plots suggest that intermediate values of S smooth the distributions better than both S = 0.10 and S = 1.00. When S = 0.10 (Figure 4.5), the smoothed distribution appear as bumpy as the unsmoothed distribution whereas when S = 1.00 (Figure 4.7), the smoothed distribution seems to overfit the unsmoothed distribution in a number of score points. A cubic spline model with S = 0.60 provides a distribution that is smooth enough and one that approximates the unsmoothed distribution at many score points along the scale. Of course, all the three models perform poorly at the higher end of the score distribution. Therefore, the cubic spline model with S = 0.60 was chosen to postsmooth the equipercentile equivalents.

4.1.2.4 Comparing Smoothing Methods

The smoothed equipercentile equivalents obtained through the use of the three smoothing methods: log-Linear model with the lower-order polynomial of degree C = 4, the four-parameter beta compound binomial model, and the cubic spline model with S = 0.60 were compared for purposes of selecting the method that would be used in the study whenever equating 2004 to 2003 form. The same criterion of identifying a model that results in a smoothed distribution that appears smooth enough without departing too much from the observed unsmoothed relationship was used. Figures 4.8 show the difference plots for the smoothed 2004 equivalents resulting from the three smoothing methods plotted on the same axes. Inspection of the graphs revealed that all the models were performing in the same way for the most part, but differed remarkably at the higher end of the score distribution. In this part of the distribution, all the three functions fell

outside plus or minus one standard error of equating (± 1SE) band. Note that even the estimation of the standard errors themselves in this region was poor due to empty cells for most of the score points. The postsmoothed cubic spline distribution (S = 0.60) and the presmoothed beta4 (beta4) distributions were performing comparatively better than the presmoothed log-Linear distribution (C = 4). Of the two, the cubic spline (S = 0.60) function was closer to the unsmoothed relationship than the beta4 function. Therefore, in this study the cubic spline model with S = 0.60 was chosen to postsmooth the distribution whenever equating 2004 to 2003 test forms.

Similar analyses were carried out to select a model for smoothing the equipercentile functions to be used when equating 2005 form to 2004 and to 2003 test forms. For these equating processes, the cubic spline with S = 1.00 was chosen because its smoothed function approximates the unsmoothed score distribution well enough.

4.1.3 Level of Motivation

The three Likert type survey questions designed to measure the degree of motivation were analyzed and the descriptive statistics and regression statistics are presented in Tables 4.6 and 4.7 respectively. Inspection of the Table 4.6 revealed that 87% of the candidates in the sample considered the test they took as important (mean = 4.65, SD = 0.85). More than 78% indicated that they prepared adequately for it (mean = 4.15, SD = 1.29) and 72% of the candidates reported that they tried hard during the course of writing (mean = 3.73, SD = 1.30). These findings imply that the vast majority of students who participated in this study were highly motivated. However, it was almost impossible, with this information alone, to conclude that the level of motivation in this study was comparable to the level of motivation on operational MANEB tests. When

performance on the test was regressed on importance, preparedness, and trying hard, the model, $Y = 23.058 + b_{import}1.198 + b_{prepare}0.191 + b_{try}0.126$, was significant as shown in Table 4.7. However, only importance seemed to be a significant predictor ($t = 2.614$, $p = 0.009$) of performance. The multiple R^2 was very small (0.009), which means that the model was explaining very little variability in performance on the tests. The squared semi-partial coefficients for all the predictors were also very small including that of the significant predictor (importance) implying that even importance was not accounting much of the performance.

4.1.4 Establishing Group Equivalency

Group equivalency was established by comparing the mean achievement of participants who took 2004 test and the mean achievement of those who took 2003 form on an external anchor test and also on the 2005 test form.

4.1.4.1 Comparing Performance on External Anchor Test

Table 4.8 shows the descriptive statistics of scores on the anchor test for both groups and Table 4.10 provides the statistics for significance testing of the group means. The mean achievement for students who took 2004 test was 16.67 whereas that for students who took 2003 form was 17.34. The skewness (0.00) and kurtosis (0.01) values in group A are close to zero which implies that scores for this group are normally distributed. The skewness (0.35) and kurtosis (0.52) values in group B are also small, which means the score distribution departs only slightly from normality. Therefore, for the data in question, the normality assumption for a t-test is met. Similarly, the Levene's test for equality of the group variances ($F = 2.98$) shows that the equal variance assumption of the t-test is satisfied ($p = 0.22$). The t-statistic for independent means ($t = -$

1.80) is not significant (df = 987, p = 0.07) at α = 0.05 and therefore, the two group means are not significantly different from each other. This result entails that the spiral procedure used during data collection was effective enough to create randomly equivalent groups.

4.1.4.2 Comparing Performance on 2005 Test

The descriptive statistics for group A and group B on the 2005 test form are presented in Table 4.9 and the statistics for significance testing of the means are presented in Table 4.10. The mean of group A on the test was 33.82 whereas the mean of group B on the same test was 34.23. The t-statistic for independent means (t = -0.539) was not significant (df = 965, p = 0.590) at α = 0.05 and therefore, the two group means were not different from each other. This finding too implies that the two groups were randomly equivalent with respect to mathematics ability and that the spiral procedure used was effective during test administration.

4.2 Is it Necessary to Equate these Tests?

4.2.1 Comparing the Difficulty of Tests

One important question that this study was investigating was whether the test forms were equally difficult. To answer this question, differences in means on the 2003 and 2004 tests for the two groups were tested using the independent t-test and Tables 4.9 and 4.10 present the results of this analysis.

Inspection of the table shows that the mean score on 2004 was 25.38 whereas the mean score on 2003 test was 33. 87. The independent t-statistic (t = -12.19, df = 987, p < 0.01) reached the significant level. The mean difference was -8.49 and the 95% confidence interval (-9.86, -7.12) for the difference showed that the difference was

statistically significant. The effect size for this difference was 0.77, which is "medium" by Cohen's (1988, cited in Crocker and Algina, 1986) criterion suggesting that the difference between the means warrants attention. Since it has been shown that the two groups of examinees were randomly equivalent, this observed difference between the group-level performances on the tests is most likely attributable to differences in the difficulty of the tests themselves. Since 2004 test had a smaller mean than 2003 form, it can be concluded that 2004 test was comparatively more difficult than 2003 test form.

4.2.2 Comparing Score Distributions

The score distributions on the two test forms were compared using log-linear analysis and by using the identity method where the equipercentile and identity equating relationships were plotted on the same axes. The relative frequency distributions for both forms are plotted in Figure 4.9 whereas in Figure 4.10 the plots of equipercentile and identity equating relationships of 2004 to 2003 scores are presented.

From the first graph, it is clear that both distributions are positively skewed, however, the 2003 distribution is shifted slightly to the right. In other words, the 2004 score distribution was more positively skewed than the 2003 distribution. In terms of kurtosis, the 2004 distribution was more leptokurtic than the 2003 score distribution. Therefore, the two distributions appear to be different. It was important to statistically investigate if this difference is significant. Following is a multinomial log-linear model that was fit to the frequencies of the test forms to test the null hypothesis that the distributions were, in fact, the same in the population of examinees:

$$\log[N_X f(x)] = w_0 + w_1 x + w_2 x^2 + w_3 x^3$$

The chi-square, $\chi^2 = 478.600$ (df = 399) is significant (p-value $\leq$ 0.004). Consequently, the null hypothesis that the two distributions are similar was rejected. According to Hanson (1992, cited in Kolen & Brennan, 2004), such a result implies that the difference between the score distributions in the population is significant and therefore, equating is necessary.

In Tables 4.11 and 4.12 the unsmoothed and smoothed conversion tables obtained from the equipercentile equating of the 2004 to the 2003 test forms using the random groups design are presented. The conversion tables for the inverse process of equating 2003 to 2004 test form is shown in Table 4.13.

Comparison of the equipercentile and the identity equating lines presented in Figure 4.10 shows that the identity equating relationship falls outside $\pm$ 2 standard errors of equipercentile equating band. Therefore, equating is considered necessary by Dorans and Lawrence's (1990) criterion.

4.3 Invariance of Examination Standards

The invariance of the examination standards on the tests were evaluated by comparing the cut scores on 2004 and 2003 tests. The grade boundaries set by the "Awards Committee" are presented in Tables 4.14 and 4.15 for 2004 and 2003 tests respectively. The operational cuts were compared to equated cuts in Table 4.16. Also presented in Table 4.17 are standard (Z) scores for each raw score point on 2004 and 2003 forms. On the PSLCE Mathematics test, candidates who obtain a score of "D" or better are certified to have passed the exam. On 2004 test, this would translate into obtaining a score of 20 or better whereas on 2003 test, it would be obtaining a score of 24 or better representing a difference of 4 raw score points. Note that a raw score of 20 on

2004 was 0.55 standard deviations below the mean of the group whereas a score of 24 on 2003 was 0.80 standard deviations below the group mean. Therefore, the cut scores for pass/fail boundaries on the two tests differ by 0.25 in standard deviation units implying that 25% fewer students met the passing criterion on 2004 form relative to 2003 form.

The cut score for a distinction category was 70 on both forms. However, this score was 4.56 standard deviations above the mean of the group on 2004 test, but only 2.94 standard deviations above the mean of the group on 2003 test. This represents a difference of 1.62 in standard deviation units. The differences between other pairs of cut scores are shown in Table 4.16. Based on these results, it would appear that the cut scores for test forms differ and the difference seems to get bigger as one move from the pass/fail grade boundary to the distinction boundary. Given that the groups were equivalent, the examination standards on these test forms were different. Since these comparisons are based on the z-scores, the difference in cut scores presented here, in itself, may be difficult interpret. Nevertheless, it would disappear if equating were instead used to maintain the same cut scores across forms.

Assuming the cuts were only set on one test form, they would easily be maintained across all subsequent forms through equating. For example, the same cuts that were set on 2003 test (reference form) would be used on 2004 test after adjusting for difficulty the scores on 2004 to a scale of scores on 2003. This scenario would mean maintaining a cut score of 24 for making a pass/fail decision on both forms. Because of the important property of moment preservation of the equating process, these cut scores would have equivalent number of standard deviations away from the group mean. In instance, equating would be instrumental in maintaining cut scores across forms.

4.4 Effect on Examinees Classification

The effect of not equating scores vis-à-vis the effect of equating test scores on examinee classification was examined by comparing the pass rates and the classification results on the 2003 and 2004 test forms before and after equating. Differences in percentages of students classified into each grade category were computed to highlight the differences in students' classification. The consistency of decisions based on 2004 and 2005 scores and also based on 2003 and 2005 scores before and after equating were also compared.

4.4.1 Pass Rates on Test Forms

Table 4.18 contains the pass rates for both forms before and after equating. Based on the cut-off points set by the committee the pass rate on 2004 form was 69.96% whereas the pass rate on 2003 form was 81.41% representing a difference of 11.45% with the pass rate for the 2004 (the more difficult test) being lower than that of 2003. Note that the groups that took 2004 and 2003 tests were randomly equivalent and as such this difference was not expected. The expectation was that the difference, attributable to sampling error, would be small and negligible. This large observed difference in cut scores, however, became smaller when equated scores were used in classifying candidates into grade boundaries.

When scores on 2004 form were equated to scores on 2003 test and when the pass/fail decisions were based on equated scores, the pass rates were 82.61% and 81.41% on 2004 and 2003 test forms respectively representing a relatively minor difference of 1.2%. Note that the pass rates that looked different before equating are not necessarily different after equating. After equating, students who were previously classified as failing

(false negatives) on 2004 test form were now classified as passing. This observation signifies the importance of equating in promoting fairness. It also helps to make pass rates across test forms more comparable to each other.

4.4.2 Classification of Candidates into Grade

Students who took 2004 and 2003 test forms were classified into different grade boundaries before and after equating and the results are presented in Table 4.19. Inspection of the Table 4.19 showed large differences in the percentage of students classified into different grade boundaries before equating. The differences, however, were largely reduced after equating. For example, in the F grade category, the difference was 11.45% before equating, but it reduced all the way to 2.37% after equating. This finding is important because it shows that maintaining cut scores across forms through equating results into more comparable classification decisions.

4.4.3 Decision Consistency

Table 4.20 shows the probabilities of making pass/fail decisions using scores on 2004 and 2005 test forms before equating whereas Table 4.21 presents probabilities of making the pass/fail decisions on the same tests after equating. The decision consistency (DC) index before equating was 0.736, which means that 73.6% of the examinees were consistently classified into pass/fail categories. After equating, the DC index was 0.694 indicating that the percent of examinees consistently classified decreased by 4 points to 69.4%. However, note that the decision consistency index used here does not take into consideration consistency of classification due to chance. The Cohen's (1988, cited in Crocker, L., Algina, J. (1986) Kappa, which considers chance consistency, was 0.252 before equating and 0.322 after equating. These values of Kappa indicate that 25.2% and

32.2% respectively of the classification decisions of students into pass/fail categories were consistent over and above chance consistency. Note that the increase in consistency above chance was high after equating and low before equating. Therefore, equating helped to increase the decision consistency. Similar results were obtained for 2003 and 2005 test forms as shown in Tables 4.22 and 4.23.

The DC before equating was 0.849 and it was 0.843 after equating. Based on the DC values, it appears that consistency of decision was equally high in both cases. However, the kappa values were 0.396 and 0.422 before and after equating respectively. Therefore, there was high consistency of classification decisions after equating (42.2%) than before equating (39.6%) indicating the importance of equating.

4.5 Equating Using External Anchor Test

The third purpose of this study was to investigate the effectiveness of using an external set of items (anchor) that were administered separately from the operational test forms to equate scores. As noted earlier, it was important in this case to establish that examinees did not change significantly in behavior of interest (mathematic ability) from the time they took the mock exam when the anchor test was administered to the time they took the 2005 test form to be equated to 2004 and 2003 forms. This in effect, implies ruling out learning effects that would confound the comparability of students' performance on the two occasions. Furthermore, it was important to establish that an anchor test administered separately from the target test can provide adequate information to be useful in equating. The external anchor test design proposed in this study would be appropriate only when these issues are established. The rest of this section reports results for the investigations into these two important issues.

4.5.1 Ruling Out Learning Effects

Students' performance were compared on different test forms in an attempt to rule out the possibility that there was much learning from the time they took the anchor test to the time they took the operational 2005 test. The mean performance of the same group of students (group A) on 2004 and 2005 test were compared through the dependent t-test as shown in Table 4.10. The group mean on 2004 was 25.51 and the group mean on 2003 test was 33.81. The paired-sample t-statistic ($t = -17.68$, $df = 481$, & $p = 0.000$) was significant indicating that group A did significantly better on 2005 form than on 2004 form. This difference could be explained either by learning effects or by differences in test difficulty.

When the group B means on 2003 and 2005 forms were compared, results were different, the group mean on 2003 test was 34.14 and mean of the same group on 2005 test was 34.23. The paired sample t-statistic ($t = -0.22$, $df = 484$, & $p = 0.829$) was not significant implying that group B performance was the same on test forms. One explanation would be that students did not learn much during the interval period or that the tests were similar in difficulty. However, these were randomly equivalent groups drawn from the same classrooms and it would not make sense to expect one random sample of students in a particular classroom to learn more than another random sample in the same classroom during the same time period. More importantly, the performance of students in groups A and B on the 2005 were already proved to be the same in earlier analyses. Given these results, differences in test difficulty, rather than learning effects, seem to account for the differences in performance for group A in that the 2004 and 2005 test forms were dissimilar in difficulty whereas the similarity in difficulty between 2004

and 2005 forms led to group B performing in the same way on both test forms. It is, therefore, conceivable to regard learning effects as minimal in this study and that the investigation into the usefulness of the external anchor test may not be confounded by this nuisance variable.

4.5.2 How Useful is the Anchor Test?

The usefulness of the anchor test was first investigated by computing the reduction in uncertainty (RIU) indices using the correlations between the scores on anchor test and scores on 2005, 2004 and 2003 tests. Secondly, results of the random groups equipercentile equating were compared to the results of the external anchor test equipercentile equating for different test forms.

4.5.2.1 Reduction in Uncertainty Index

The RIU indices for test forms are presented in Table 4.24. The indices are small for all tests signifying that the anchor test scores were reducing the uncertainty about the test forms only to a small degree. For example, the anchor test reduces the uncertainty of knowing scores on 2004, and 2003 test forms by 14% and 17% respectively. Therefore, significant amount of uncertainty about 2005, 2004, and 2003 tests still remain after including information from the anchor test. This suggests that the anchor test that was developed would not serve as a useful variable during equating. Nevertheless, note that the RIU index for 2005 test was as low as the indices for 2004 and 2003 tests. This is an important observation because it entails that administering the anchor test 5 weeks away from 2005 test did not have a different impact on the uncertainty about the test from what would be known if the anchor test was administered two days away from the test as was the case with 2004 and 2003 forms.

4.5.2.2 Comparing Equating Designs

The random groups equipercentile equating function of 2004 test to 2003 form were compared to the external anchor test equipercentile equating function of 2004 test to 2003 form. Similar comparisons were made between the random groups equipercentile equating functions and their corresponding external anchor test equipercentile functions for the 2005 test to 2004 form and the 2005 test to 2003 test. The comparisons were facilitated by computing the average root mean square differences (RMSD) shown in Table 4.25 and through plots shown in Figures 4.11, 4.12, and 4.13.

The results in Table 4.25 indicate that the difference between the random groups equipercentile equating and the external anchor test equipercentile relationships in all cases were small. The average RMSD for 2004 to 2003 equating functions was 0.032, the average RMSD foe the 2005 to 2004 equating functions was 0.171, and for the 2005 and 2003 equating functions, the average RMDS was 0.063. The standardized root mean square differences presented in the Table 4.25 could also be regarded as effect size measures (Doran, 2000). In this context, the effect of the difference seems to be very small and negligible. This finding is important in that it establishes the fact that, in spite of the small RIU indices, the external anchor test design is as useful, in this study, as the random groups design.

Finally, the equated functions obtained from the random groups design and from the external anchor test design were compared by plotting them on the same axes as shown in Figure 4.11 for the 2004 test to 2003 test, in Figure 4.12 for the 2005 test to 2004 test, and in Figure 4.13 for the 2005 test to 2003 test form. In all cases, the plots indicate that there are very small differences between the equating functions for the most

part of the score scale. Of course, the lines do not necessary superimpose on each other in certain scores points along the score scale, but they are close to each other for the most part of the scale. The differences were comparatively smaller for 2004 test to 2003 test equating functions (Figure 4.11) and for the 2005 test to 2003 test equating functions (Figure 4.13) than the difference for the 2005 test to 2004 test equating functions (Figure 4.12). Once again, this important result confirms earlier findings that there is no difference in the way the two equating designs perform in this study. Therefore, it is reasonable to proceed with equating 2005 test to the 2004 form and also equating 2005 test to 2003 form using the external anchor test.

4.5.3 Equating via the Anchor Test

4.5.3.1 Conversion Tables

The scores on the 2005 test form were equated to scores on 2004 and 2003 test forms using equipercentile frequency estimation method. Table 4.26 shows the resulting conversion table for 2005 test to the 2004 test whereas Table 4.27 presents the conversion table for the 2005 test to the 2003 test form. Later the 2005 test was also equated to the scale of 2004 and 2003 test forms using the Tucker linear equating. The resulting conversion tables are given in Table 4.28 (for 2005 to 2004 form) and in Table 4.29 (for 2005 test to 2003 form).

4.5.3.2 Mean Square Equating Errors

The mean square equating errors (MSEE) shown in Table 4.25 were also computed to assess the adequacy of both linear and equipercentile equating processes. The MSEE resulting from the equipercentile equating of 2005 to the 2004 test forms was 2.76 and that from the equipercentile equating of the 2005 to the 2003 test forms was

1.43. These are generally small equating errors in themselves. The MSEE resulting from the Tucker linear equating of 2005 to the 2004 test forms was 3.95 and that from the Tucker linear equating of the 2005 to the 2003 test forms was 3.59. These too are generally small equating errors in themselves, but they are comparatively larger than the equipercentile equating errors. Therefore, the conclusion based on the magnitude of the MSEE is that the equating processes were reasonably adequate.

4.5.3.3 Comparing Students' Classification

Table 4.30 presents the classification of students into pass/fail categories on 2005, 2004 and 2003 test forms before and after equipercentile equating. Before equating, the pass rate on 2005 was 89.69% and it was 70.33% on 2004 test form representing a difference of 19.36%. After equating the pass rate on 2005 test form changed to 69.07% and the difference decreased to 1.26%. This finding suggests that equating through the external anchor test was better than not equating at all in supporting fair classification decisions. Similarly, equating helped to reduce the gap in pass rate between 2005 and 2003 test forms. Before equating the pass rate on 2005 was 86.28% whereas on 2003 form the pass rate was 81.44% representing a difference of 4.84%. After equipercentile equating, the difference decreased to 1.19%. Therefore, this finding too signifying that equating through the external anchor test was better than not equating at all.

Table 4.31 presents the classification of students into pass/fail categories on 2005, 2004 and 2003 test forms before and after the Tucker linear equating. As before, the difference in pass rate on 2005 test and on 2004 test was 19.36% before equating. However, after linear equating the pass rate on 2005 test form changed to 73.40% and it was 70.33% on 2003 form representing a difference of 3.07%. Therefore, equating

through the external anchor test resulted into fairer classification decisions than not equating. However, linear equating did not reduce the gap in pass rate between 2005 and 2003 test forms. Before equating, the difference in pass rate was 4.84% and after equating, the difference was 4.31%. Except for this one result, the findings presented in this section signifying that equating through the external anchor test was better than not equating at all.

4.6 Summary of Results

The preliminary results for the study show that the reliability of the test forms was low indicating that the items in 2004 and 2003 forms were not consistently working together in the two tests. The external anchor test items that were apparently mirrored after the items in test forms also had low reliability. It was less surprising, therefore, that the anchor test was not highly correlated with the test forms. On the positive note, the participants in the study were highly motivated and finally, based on their performance on external anchor and 2005 tests, the two randomly selected groups were equivalent with respect to the mathematics ability.

On the main data analyses, the results have shown that the mean scores for the two groups on the 2004 and the 2003 test forms were significantly different. This difference was attributed to differences in difficulty of the test forms themselves since the groups were randomly equivalent. The distributions of scores, too, for the two groups were different, which was indicative that it was necessary to equate scores on the test forms to adjust scores for difficulty and to match distributions of scores. Furthermore, comparing the equipercentile and the identity equating lines revealed that equating was necessary.

The results also showed that cut scores on the test forms varied across years. Because the cut scores were not invariant, the pass rates were different and students were classified differently into the grade boundaries. After equating scores on the test forms, the difference between the pass rates vanished and the differences in the way candidates were being classified also decreased.

The results further reveal that although the external anchor test was only moderately correlated with the test forms, equating scores through the anchor produced results that were not different from those obtained using the random groups design. Because of this finding, the study presented conversion tables for mapping the 2005 to the 2004 test forms and others for mapping the 2005 to the 2003 test forms.

Table 4.1: Item-Total Correlation, Mean and Alpha Values for 2004 and 2003 Tests

	2004 Test Form		2003 Test Form	
Item	Item-Total Correlation	Mean (p-value)	Item-Total Correlation	Mean (p-value)
1	.02	.22	.19	.77
2	.21	.73	-.01	.14
3	.16	.63	.22	.46
4	.10	.21	.23	.44
5	.21	.34	.24	.43
6	.25	.40	.29	.34
7	.10	.25	.23	.30
8	.10	.22	-.04	.34
9	.08	.39	.15	.63
10	.29	.26	.02	.24
11	.18	.33	.37	.42
12	.11	.55	.03	.16
13	.08	.13	.08	.17
14	.21	.51	.05	.21
15	.20	.33	.23	.38
16	.28	.32	.21	.43
17	.11	.16	.27	.51
18	.06	.07	.17	.40
19	.10	.20	.24	.30
20	.21	.30	.07	.37
21	.00	.10	.19	.35
22	.04	.18	.23	.71
23	.12	.26	-.05	.25
24	.02	.18	.28	.66
25	.03	.17	.36	.57
26	.21	.33	-.03	.19
27	.04	.22	.06	.36
28	.12	.77	.15	.62
29	.20	.66	.07	.31
30	.24	.21	.14	.35
31	.29	.50	.26	.23
32	.34	.13	.25	.11
33	.15	.07	.39	.48
34	.21	.03	.32	.18
35	.12	.10	.37	.15
$\bar{X}_i$	.15	.30	.18	.37
S_i	.08	.19	.13	.17
α (2004) = .63			α (2003) = .67	

Table 4.2: Item-Total Correlation, Mean and Alpha Values for Anchor Test

Item	Item-Total Correlation	Item Mean
1	.20	.85
2	.14	.69
3	.24	.30
4	.11	.08
5	.09	.18
6	.22	.37
7	.22	.50
8	.08	.13
9	.13	.44
10	.26	.19
11	.22	.61
$\bar{X}_i$	.17	.39
S_i	.07	.25
$\alpha = .49$		

Table 4.3: Correlation Coefficients Between Tests and subtests

Test Pair	r	r^2
Anchor vs 2003	.561	.315
Anchor vs 2004	.511	.261
Anchor vs 2005 (P1)	.515	.265
Anchor vs 2005 (P2)	.506	.256
Anchor vs 2005 (Total)	.509	.259
MC vs CR (2003)	.532	.283
MC vs CR (2004)	.380	.144

Table 4.4: Moments for Presmoothing Score Distributions

Form/Method	Mean($\hat{\mu}$)	SD($\hat{\sigma}$)	Skewness($\bar{sk}$)	Kurtosis($\bar{ku}$)
2003				
Method				
Unsmoothed	33.867	12.294	0.725	4.287
Beta4	33.867	12.294	0.725	3.310
Log-Linear				
C = 10	33.867	12.294	0.725	4.287
C = 9	33.867	12.294	0.725	4.287
C = 8	33.867	12.294	0.725	4.287
C = 7	33.867	12.294	0.725	4.287
C = 6	33.867	12.294	0.725	4.287
C = 5	33.867	12.294	0.725	4.287
C = 4	33.867	12.294	0.725	4.287
C = 3	33.867	12.294	0.725	4.606
C = 2	33.867	12.294	0.063	2.875
C = 1	33.867	26.414	0.695	2.481
2004				
Method				
Unsmoothed	25.376	9.769	1.080	5.754
Beta4	25.375	9.769	1.080	4.428
Log-Linear				
C = 10	25.376	9.769	1.080	5.754
C = 9	25.376	9.769	1.080	5.754
C = 8	25.376	9.769	1.080	5.754
C = 7	25.376	9.769	1.080	5.754
C = 6	25.376	9.769	1.080	5.754
C = 5	25.376	9.769	1.080	5.756
C = 4	25.376	9.769	1.080	5.754
C = 3	25.376	9.769	1.080	8.202
C = 2	25.376	9.769	0.087	2.845
C = 1	25.376	22.566	1.129	3.658

Table 4.5: Moments for Postsmoothed Score Distributions

Test Form	Mean($\hat{\mu}$)	SD($\hat{\sigma}$)	Skewness($\hat{sk}$)	Kurtosis($\hat{ku}$)
2003	33.867	12.294	0.725	4.287
2004	25.376	9.769	1.080	5.754
2004 Equated to 2003 Scale				
Unsmoothed	33.868	12.272	0.724	4.276
S = 0.10	33.840	12.172	0.644	3.843
S = 0.20	33.844	12.175	0.656	3.908
S = 0.30	33.846	12.173	0.663	3.939
S = 0.40	33.853	12.185	0.668	3.934
S = 0.50	33.858	12.188	0.668	3.922
S = 0.60	33.856	12.193	0.668	3.920
S = 0.70	33.844	12.188	0.677	3.943
S = 0.80	33.822	12.177	0.692	3.985
S =1.00	33.749	12.147	0.736	4.095

Table 4.6: Descriptive Statistics for Survey Questions

		Percent of Students in Each Rating Category					Summary Statistics		
Item	N	1	2	3	4	5	$\bar{X}_i$	S_i	Median
Importance	1006	1.60	2.30	8.00	6.10	81.00	4.65	.85	5
Preparedness	1005	7.40	8.80	5.00	18.30	60.00	4.15	1.29	5
Tried Hard	1004	9.40	13.20	4.00	40.50	31.60	3.73	1.30	4

Note: Importance: 1 = Not at all Important to 5 = Very Important; Preparedness: 1 = Not at all Prepared to 5 = Very Much Prepared; and Tried Hard: 1 = Never Tried Hard to 5 = A Great Deal Hard.

Table 4.7: Regression Statistics for Survey Questions

	Regression Statistics							
	R	R^2	b_0	b_1	t	p	$r_{y(i.jk)}$	$r^2_{y(i.jk)}$
Model	.095	.009	23.058		9.817	.000		
Importance				1.198	2.614	.009	.084	.007
Preparedness				.191	.619	.536	.020	.000
Tried Hard				.126	.411	.681	.013	.000

Table 4.8: Descriptive Statistics of Scores on Anchor Test.

	GROUP		TOTAL SAMPLE
DESCRIPTIVE STATISTIC	A	B	
N	489	500	989
Mean ($\bar{X}_i$)	16.67	17.34	17.01
Standard Deviation (S_i)	5.56	6.06	5.06
SE of Mean	.25	.27	.19
Skewness	.00	.35	.22
Kurtosis	.01	.52	.37

Table 4.9: Descriptive Statistics of Scores on Test Forms

STATISTIC	TEST FORM				
	2004	2003	2005		
			A(2004)	B(2003)	TOTAL
N	506	511	482	485	967
Mean	25.38	33.87	33.82	34.23	34.02
SD	9.78	12.31	12.11	11.65	11.88
SE of $\bar{X}_t$	.44	.54	.55	.53	.38
Skewness	1.08	.73	.58	.91	.73
Kurtosis	2.77	1.31	.51	1.54	.99

Table 4.10: Group Differences on Tests Forms

ANALYSIS	$Diff_{\bar{X}}$	t	df	p	95%CI		Δ
					Lower	Upper	
A vs B On Anchor	-.67	-1.80	987	.07	-1.39,	.05	.12
A(2004) Vs B(2003)	-8.50	-12.19	970	.00	-9.86,	-7.12	.77
A vs B On 2005	-.41	-.54	965	.59	-1.91,	1.09	.03
A(2004) Vs A(2005)	-8.31	-17.68	481	.00	-9.23,	-7.38	.75
B(2003) Vs B(2005)	-.36	-.22	484	.83	-.91,	.73	.03

Table 4.11: Unsmoothed Raw-to- Raw Score Conversion Table for 2004 to 2003 Scores.

2004 Score	2003 Equivalents	2004 Score	2003 Equivalents	2004 Score	2003 Equivalents
0	0.000	34	45.169	68	76.480
1	1.000	35	46.796	69	76.480
2	2.000	36	48.040	70	76.480
3	3.000	37	49.359	71	76.480
4	4.000	38	50.652	72	76.480
5	5.000	39	51.722	73	76.480
6	7.668	40	53.168	74	81.985
7	7.837	41	54.635	75	82.490
8	8.173	42	54.972	76	82.490
9	9.681	43	55.309	77	82.490
10	11.797	44	56.406	78	92.995
11	14.026	45	60.784	79	93.500
12	16.530	46	61.941	80	93.500
13	17.698	47	62.446	81	93.500
14	18.648	48	64.203	82	93.500
15	19.938	49	64.916	83	93.500
16	21.612	50	65.421	84	93.500
17	23.697	51	65.421	85	93.500
18	25.512	52	65.421	86	93.500
19	26.792	53	65.421	87	93.500
20	27.785	54	65.421	88	93.500
21	28.826	55	67.431	89	93.500
22	29.643	56	68.318	90	93.500
23	31.072	57	72.470	91	93.500
24	32.528	58	72.470	92	93.500
25	34.100	59	72.470	93	93.500
26	35.126	60	72.470	94	94.000
27	36.003	61	72.470	95	95.000
28	37.058	62	72.470	96	96.000
29	38.567	63	72.470	97	97.000
30	39.504	64	75.975	98	98.000
31	40.293	65	76.480	99	99.000
32	41.270	66	76.480	100	100.000
33	43.304	67	76.480		

Table 4.12: Smoothed Raw-to-Raw Score Conversion Table for 2004 to 2003 Scores

2004 Score	2003 Equivalents	2004 Score	2003 Equivalent	2004 Score	2003 Equivalent
0	0.10	34	45.04	68	79.39
1	1.29	35	46.33	69	80.04
2	2.48	36	47.60	70	80.69
3	3.68	37	48.87	71	81.34
4	4.87	38	50.13	72	81.99
5	6.06	39	51.37	73	82.64
6	7.25	40	52.60	74	83.29
7	8.45	41	53.81	75	83.94
8	9.64	42	55.00	76	84.59
9	11.12	43	56.16	77	85.24
10	12.74	44	57.31	78	85.89
11	14.34	45	58.43	79	86.54
12	15.91	46	59.53	80	87.19
13	17.46	47	60.62	81	87.84
14	18.97	48	61.68	82	88.48
15	20.44	49	62.72	83	89.13
16	21.90	50	63.75	84	89.78
17	23.32	51	64.76	85	90.43
18	24.71	52	65.77	86	91.08
19	26.07	53	66.76	87	91.73
20	27.40	54	67.75	88	92.38
21	28.70	55	68.74	89	93.03
22	29.99	56	69.72	90	93.68
23	31.25	57	70.70	91	94.33
24	32.51	58	71.67	92	94.98
25	33.76	59	72.64	93	95.63
26	35.00	60	73.61	94	96.28
27	36.24	61	74.56	95	96.93
28	37.48	62	75.35	96	97.58
29	38.72	63	76.14	97	98.23
30	39.97	64	76.79	98	98.88
31	41.22	65	77.44	99	99.53
32	42.49	66	78.09	100	100.18
33	43.76	67	78.74		

Table 4.13: Smoothed Raw-to-Raw Score Conversion Table for 2003 to 2004 Scores.

2003 Score	2004 Equivalents	2003 Score	2004 Equivalent	2003 Score	2004 Equivalent
0	-0.08	34	25.19	68	54.26
1	0.77	35	26.00	69	55.27
2	1.61	36	26.81	70	56.30
3	2.45	37	27.62	71	57.32
4	3.30	38	28.42	72	58.35
5	4.14	39	29.23	73	59.38
6	4.98	40	30.03	74	60.42
7	5.83	41	30.82	75	61.45
8	6.67	42	31.61	76	62.78
9	7.52	43	32.40	77	64.32
10	8.18	44	33.18	78	65.86
11	8.85	45	33.97	79	67.40
12	9.51	46	34.75	80	68.94
13	10.14	47	35.53	81	70.48
14	10.77	48	36.31	82	72.01
15	11.41	49	37.10	83	73.55
16	12.05	50	37.90	84	75.09
17	12.70	51	38.70	85	76.63
18	13.36	52	39.51	86	78.17
19	14.03	53	40.33	87	79.71
20	14.70	54	41.17	88	81.25
21	15.38	55	42.01	89	82.79
22	16.08	56	42.87	90	84.33
23	16.78	57	43.74	91	85.87
24	17.49	58	44.62	92	87.41
25	18.21	59	45.52	93	88.95
26	18.95	60	46.43	94	90.49
27	19.70	61	47.36	95	92.03
28	20.46	62	48.31	96	93.57
29	21.23	63	49.27	97	95.11
30	22.01	64	50.25	98	96.65
31	22.80	65	51.24	99	98.19
32	23.59	66	52.24	100	99.73
33	24.39	67	53.24		

Table 4.14: Grade Boundaries for 2004 Test Form

	GRADE BOUNDARIES		NUMBER AND % OF CANDIDATES IN EACH GRADE		CUMMULATIVE %
GRADE	FROM	TO	NUMBER	%	%
A	70	100	2	0.40	0.40
B	55	69	6	1.18	1.58
C	33	54	99	19.57	21.15
D	20	32	247	48.81	69.96
F	0	19	152	30.04	100.00

Table 4.15: Grade Boundaries for 2003 Test Form

	GRADE BOUNDARIES		NUMBER AND % OF CANDIDATES IN EACH GRADE		CUMMULATIVE %
GRADE	FROM	TO	NUMBER	%	%
A	70	100	4	0.78	0.78
B	56	69	17	3.33	4.11
C	35	55	209	40.90	45.01
D	24	34	186	36.40	81.41
F	0	23	95	18.59	100.00

Table 4.16: Operational and Equated Cut Scores

	Operational Cuts		Absolute Difference		Equated Cuts	
Grade Boundaries	2004	2003	Raw Score	SD Units	2004	2003
B/A	70	70	0.00	1.63	81	56
C/B	55	56	1.00	1.23	69	43
D/C	33	34	1.00	0.77	44	26
F/D	20	24	4.00	0.25	27	17

Table 4.17: Standard Score (Z-Score)

Score	Z_{2004}	Z_{2003}	Score	Z_{2004}	Z_{2003}	Score	Z_{2004}	Z_{2003}
0	-2.595	-2.752	34	0.881	0.011	68	4.358	2.773
1	-2.493	-2.671	35	0.984	0.092	69	4.461	2.855
2	-2.391	-2.590	36	1.086	0.173	70	4.563	2.936
3	-2.289	-2.509	37	1.188	0.254	71	4.665	3.017
4	-2.186	-2.427	38	1.291	0.336	72	4.767	3.098
5	-2.089	-2.346	39	1.393	0.417	73	4.870	3.180
6	-1.982	-2.265	40	1.495	0.498	74	4.972	3.261
7	-1880	-2.183	41	1.597	0.579	75	5.074	3.342
8	-1.777	-2.102	42	1.700	0.661	76	5.176	3.424
9	-1.675	-2.021	43	1.802	0.742	77	5.279	3.505
10	-1.573	-1.940	44	1.904	0.823	78	5.381	3.586
11	-1.470	-1.858	45	2.006	0.904	79	5.483	3.667
12	-1.368	-1.777	46	2.109	0.986	80	5.585	3.749
13	-1.266	-1.696	47	2.211	1.067	81	5.688	3.830
14	-1.164	-1.615	48	2.313	1.148	82	5.790	3.911
15	-1.061	-1.533	49	2.415	1.229	83	5.892	3.992
16	-0.959	-1.452	50	2.518	1.311	84	5.994	4.074
17	-0.857	-1.371	51	2.620	1.392	85	6.097	4.155
18	-0.755	-1.290	52	2.722	1.473	86	6.199	4.236
19	-0.652	-1.208	53	2.824	1.555	87	6.301	4.317
20	-0.550	-1.127	54	2.927	1.636	88	6.404	4.399
21	-0.448	-1.046	55	3.029	1.717	89	6.506	4.480
22	-0.346	-0.965	56	3.131	1.798	90	6.608	4.561
23	-0.243	-0.883	57	3.233	1.880	91	6.710	4.642
24	-0.141	-0.802	58	3.336	1.961	92	6.813	4.724
25	-0.039	-0.721	59	3.438	2.042	93	6.915	4.805
26	0.063	-0.640	60	3.540	2.123	94	7.017	4.886
27	0.166	-0.558	61	3.642	2.205	95	7.119	4.967
28	0.268	-0.477	62	3.745	2.286	96	7.222	5.049
29	0.370	-0.396	63	3.847	2.367	97	7.324	5.130
30	0.472	-0.314	64	3.949	2.448	98	7.426	5.211
31	0.575	-0.233	65	4.052	2.530	99	7.528	5.293
32	0.677	-0.152	66	4.154	2.611	100	7.631	5.374
33	0.779	-0.071	67	4.256	2.692			

Table 4.18: Pass Rates on 2004 and 2003 Test Forms

Form	N	Pass	Fail	Pass Rate	Difference
Not Equated Scores					
2004	506	354	152	69.96%	
					11.45%
2003	511	416	75	81.41%	
Equated Scores					
2004	506	418	88	82.61%	
					1.20%
2003	511	416	75	81.41%	

Table 4.19: Classification of Candidates Using Cut Scores on the Reference Form (2003)

	Scores Not Equated				Scores Equated			
Grade/ Form	Cuts	N	%	Absolute Difference (%)	Cuts	N	%	Absolute Difference %)
A								
2004	70	2	0.40		70	3	0.59	
				0.38				0.19
2003	70	4	0.78		70	4	0.78	
B								
2004	55	6	1.18		56	22	4.35	
				2.12				1.02
2003	56	17	3.33		56	6	3.33	
C								
2004	33	99	19.57		35	198	39.13	
				21.33				1.77
2003	35	209	40.90		35	209	40.90	
D								
2004	20	247	48.81		24	177	34.98	
				12.41				1.42
2003	24	186	36.40		24	186	36.40	
F								
2004	0	152	30.04		0	106	20.95	
				11.45				2.37
2003	0	95	18.59		0	95	18.58	

Table 4.20: Decision Consistency Analysis for 2005 and 2004 Tests before Equating

		Operational 2005 Scores: Pass	Operational 2005 Scores: Fail	
2004 Scores	Pass	N = 313 $P_{00} = 0.650728$	N = 25 $P_{01} = 0.051975$	$P_{0.} = 0.702703$
	Fail	N = 102 $P_{10} = 0.212058$	N = 41 $P_{11} = 0.085239$	$P_{1.} = 0.297297$
		$P_{.0} = 0.862786$	$P_{.1} = 0.137214$	

Decision Consistency (P) = 0.650728 + 0.085239 = 0.735967

Chance Consistency (P_C) = $P_{1.}P_{.1} + P_{0.}P_{.0}$ = 0.647075

Cohen's Kappa (κ) = $\frac{P - P_C}{1 - P_C}$ = 0.251871

Table 4.21: Decision Consistency Analysis for 2005 and 2004 Tests after Equating

		Equated 2005 Scores: Pass	Equated 2005 Scores: Fail	
2004 Scores	Pass	N = 245 $P_{00} = 0.509356$	N = 93 $P_{01} = 0.193347$	$P_{0.} = 0.702703$
	Fail	N = 54 $P_{10} = 0.112266$	N = 89 $P_{11} = 0.185031$	$P_{1.} = 0.297297$
		$P_{.0} = 0.621622$	$P_{.1} = 0.378378$	

Decision Consistency (P) = 0.509356 + 0.185031 = 0.694387

Chance Consistency (P_C) = $P_{1.}P_{.1} + P_{0.}P_{.0}$ = 0.549306

Cohen's Kappa (κ) = $\frac{P - P_C}{1 - P_C}$ = 0.321905

Table 4.22: Decision Consistency Analysis for 2005 and 2003 Tests before Equating

	Operational 2005 Scores: Pass	Operational 2005 Scores: Fail	
2003 Scores: Pass	N = 379 $P_{00} = 0.781443$	N = 16 $P_{01} = 0.03299$	$P_{0.} = 0.814433$
2003 Scores: Fail	N = 57 $P_{10} = 0.117526$	N = 33 $P_{11} = 0.068041$	$P_{1.} = 0.185567$
	$P_{.0} = 0.898969$	$P_{.1} = 0.101031$	

Decision Consistency (P) = 0.781443 + 0.068041 = 0.849485

Chance Consistency (P_C) = $P_{1.}P_{.1} + P_{0.}P_{.0}$ = 0.750898

Cohen's Kappa (κ) = $\dfrac{P - P_C}{1 - P_C}$ = 0.395768

Table 4.23: Decision Consistency Analysis for 2005 and 2003 Tests after Equating

	Equated 2005 Scores: Pass	Equated 2005 Scores: Fail	
2003 Scores: Pass	N = 369 $P_{00} = 0.760825$	N = 26 $P_{01} = 0.053608$	$P_{0.} = 0.814433$
2003 Scores: Fail	N = 50 $P_{10} = 0.103098$	N = 40 $P_{11} = 0.082474$	$P_{1.} = 0.185567$
	$P_{.0} = 0.863918$	$P_{.1} = 0.136082$	

Decision Consistency (P) = 0.760825 + 0.082474 = 0.843299

Chance Consistency (P_C) = $P_{1.}P_{.1} + P_{0.}P_{.0}$ = 0.728855

Cohen's Kappa (κ) = $\dfrac{P - P_C}{1 - P_C}$ = 0.42207

Table 4.24: Reduction in Uncertainty Indices (RIU)

Test	$\hat{\mu}$	$\hat{\sigma}$	r	r^2	RIU
2004	25.510	9.881	0.511	0.261	0.140
2003	34.140	12.415	0.561	0.315	0.172
2005 (GA)	33.820	12.108	0.515	0.265	0.143
2005 (GB)	34.230	11.647	0.506	0.256	0.137

Note: GA represents group A of examinees whereas GB represents group B of examinees.

Table 4.25: Standardized Root Mean Square Differences (RMSDs) and Mean Square Equating Errors (MSEE)

Test Forms Equated	RG	AT	TL	RG vs. AT
	MSEE	MSEE	MSEE	RMSD
2004 to 2003	2.523	2.117	9.107	0.032
2005 to 2004	1.496	2.761	3.953	0.171
2005 to 2003	2.211	1.438	3.589	0.063

NOTE: RG = Random Groups Equipercentile Equating; AT = Anchor Test Equipercentile Equating; and TL = Tucker Linear Equating.

Table 4.26: Equipercentile Raw-to-Raw Score Conversion Table for 2005 to 2004 Scores

2005 Score	2004 Equivalents	2005 Score	2004 Equivalent	2005 Score	2004 Equivalent
0	0.06	34	25.91	68	56.67
1	0.67	35	26.83	69	57.55
2	1.28	36	27.75	70	58.43
3	1.90	37	28.67	71	59.31
4	2.51	38	29.58	72	60.20
5	3.13	39	30.50	73	61.09
6	3.75	40	31.40	74	61.99
7	4.35	41	32.31	75	62.89
8	4.97	42	33.21	76	63.80
9	5.58	43	34.10	77	64.99
10	6.19	44	34.98	78	66.19
11	6.81	45	35.86	79	67.39
12	7.40	46	36.73	80	68.59
13	7.99	47	37.60	81	69.79
14	8.59	48	38.46	82	71.36
15	9.33	49	39.33	83	72.94
16	10.12	50	40.20	84	74.51
17	10.91	51	41.07	85	76.09
18	11.72	52	41.96	86	77.66
19	12.54	53	42.85	87	79.24
20	13.37	54	43.75	88	80.81
21	14.21	55	44.67	89	82.39
22	15.06	56	45.59	90	83.96
23	15.93	57	46.52	91	85.54
24	16.80	58	47.46	92	87.11
25	17.69	59	48.40	93	88.69
26	18.59	60	49.35	94	90.26
27	19.49	61	50.30	95	91.84
28	20.40	62	51.24	96	93.41
29	21.32	63	52.17	97	94.99
30	22.24	64	53.09	98	96.56
31	23.15	65	54.00	99	98.14
32	24.07	66	54.90	100	99.71
33	24.99	67	55.79		

Table 4.27: Equipercentile Raw-to-Raw Score Conversion Table for 2005 to 2003 Scores

2005 Score	2003 Equivalents	2005 Score	2003 Equivalent	2005 Score	2003 Equivalent
0	0.20	34	36.63	68	66.34
1	1.10	35	34.62	69	67.34
2	1.99	36	35.60	70	68.33
3	2.89	37	36.57	71	69.33
4	3.79	38	37.54	72	70.42
5	4.68	39	38.51	73	71.50
6	5.58	40	39.47	74	72.59
7	6.48	41	40.43	75	73.68
8	7.38	42	41.39	76	74.76
9	8.27	43	42.35	77	75.82
10	9.26	44	43.30	78	76.87
11	10.28	45	44.25	79	77.92
12	11.31	46	45.20	80	78.97
13	12.33	47	46.14	81	80.02
14	13.36	48	47.09	82	81.07
15	14.38	49	48.04	83	82.12
16	15.41	50	48.98	84	83.17
17	16.43	51	49.93	85	84.22
18	17.46	52	50.87	86	85.27
19	18.48	53	51.82	87	86.32
20	19.51	54	52.77	88	87.37
21	20.53	55	53.72	89	88.42
22	21.56	56	54.67	90	89.47
23	22.58	57	55.63	91	90.52
24	23.60	58	56.59	92	91.57
25	24.62	59	57.55	93	92.62
26	25.63	60	58.51	94	93.67
27	26.64	61	59.48	95	94.72
28	27.65	62	60.45	96	95.77
29	28.66	63	61.42	97	96.82
30	29.66	64	62.40	98	97.87
31	30.66	65	63.38	99	98.92
32	31.65	66	64.36	100	99.97
33	32.64	67	65.35		

Table 4.28: Linear Raw-to-Raw Score Conversion Table for 2005 to 2004 Scores

2005 Score	2004 Equivalents	2005 Score	2004 Equivalent	2005 Score	2004 Equivalent
0	-3.61	34	25.94	68	55.49
1	-2.74	35	26.81	69	56.36
2	-1.87	36	27.68	70	57.22
3	-1.00	37	28.55	71	58.09
4	-0.13	38	29.41	72	58.96
5	0.74	39	30.28	73	59.83
6	1.60	40	31.15	74	60.70
7	2.47	41	32.02	75	61.57
8	3.34	42	32.89	76	62.44
9	4.21	43	33.76	77	63.31
10	5.08	44	34.63	78	64.18
11	5.95	45	35.50	79	65.05
12	6.82	46	36.37	80	65.92
13	7.69	47	37.24	81	66.78
14	8.56	48	38.11	82	67.65
15	9.43	49	38.97	83	68.52
16	10.30	50	39.82	84	69.39
17	11.16	51	40.71	85	70.26
18	12.03	52	41.58	86	71.13
19	12.90	53	42.45	87	72.00
20	13.77	54	43.32	88	72.87
21	14.64	55	44.19	89	73.74
22	15.51	56	45.06	90	74.61
23	16.38	57	45.93	91	75.47
24	17.25	58	46.80	92	76.34
25	18.12	59	47.66	93	77.21
26	18.99	60	48.53	94	78.08
27	19.85	61	49.40	95	78.95
28	20.72	62	50.27	96	79.82
29	21.59	63	51.14	97	80.69
30	22.46	64	52.01	98	81.56
31	23.33	65	52.88	99	82.43
32	24.20	66	53.75	100	83.30
33	25.07	67	54.62		

Table 4.29: Linear Raw-to-Raw Score Conversion Table for 2005 to 2003 Scores

2005 Score	2003 Equivalents	2005 Score	2003 Equivalent	2005 Score	2003 Equivalent
0	-0.79	34	36.46	68	67.70
1	0.22	35	34.46	69	68.71
2	1.22	36	35.47	70	69.72
3	2.23	37	36.48	71	70.72
4	3.24	38	37.48	72	71.73
5	4.25	39	38.49	73	72.74
6	5.25	40	39.50	74	73.75
7	6.26	41	40.51	75	74.75
8	7.27	42	41.51	76	75.76
9	8.27	43	42.52	77	76.77
10	9.28	44	43.53	78	77.77
11	10.29	45	44.53	79	78.78
12	11.30	46	45.54	80	79.75
13	12.30	47	46.55	81	80.80
14	13.31	48	47.56	82	81.80
15	14.32	49	48.56	83	82.81
16	15.32	50	49.57	84	83.82
17	16.33	51	50.58	85	84.82
18	17.34	52	51.59	86	85.83
19	18.35	53	52.59	87	86.84
20	19.35	54	53.60	88	87.85
21	20.36	55	54.61	89	88.85
22	21.37	56	55.61	90	89.86
23	22.38	57	56.62	91	90.87
24	23.38	58	57.63	92	91.88
25	24.39	59	58.64	93	92.88
26	25.40	60	59.64	94	93.89
27	26.40	61	60.65	95	94.90
28	27.41	62	61.66	96	95.90
29	28.42	63	62.67	97	96.91
30	29.43	64	63.67	98	97.92
31	30.43	65	64.68	99	98.93
32	31.44	66	65.69	100	99.93
33	32.45	67	66.69		

Table 4.30: Pass Rates on 2005, 2004 and 2003 Tests before and after Equipercentile Equating.

Form	N	Pass	Fail	Pass Rate	Absolute Difference
Before Equating					
2005	485	435	50	89.69%	
					19.36%
2004	482	339	143	70.33%	
After Equating					
2005	485	335	150	69.07%	
					1.26%
2004	482	339	143	70.33%	
Before Equating					
2005	481	415	66	86.28%	
					4.84%
2003	485	395	90	81.44%	
After Equating					
2004	481	386	95	80.25%	
					1.19%
2003	485	395	90	81.44%	

Table 4.31: Pass Rates on 2005, 2004 and 2003 Tests before and after Tucker Linear Equating.

Form	N	Pass	Fail	Pass Rate	Absolute Difference
Before Equating					
2005	485	435	50	89.69%	
					19.36%
2004	482	339	143	70.33%	
After Equating					
2005	485	356	129	73.40%	
					3.07%
2004	482	339	143	70.33%	
Before Equating					
2005	481	415	66	86.28%	
					4.84%
2003	485	395	90	81.44%	
After Equating					
2005	481	371	110	77.13%	
					4.31%
2003	485	395	90	81.44%	

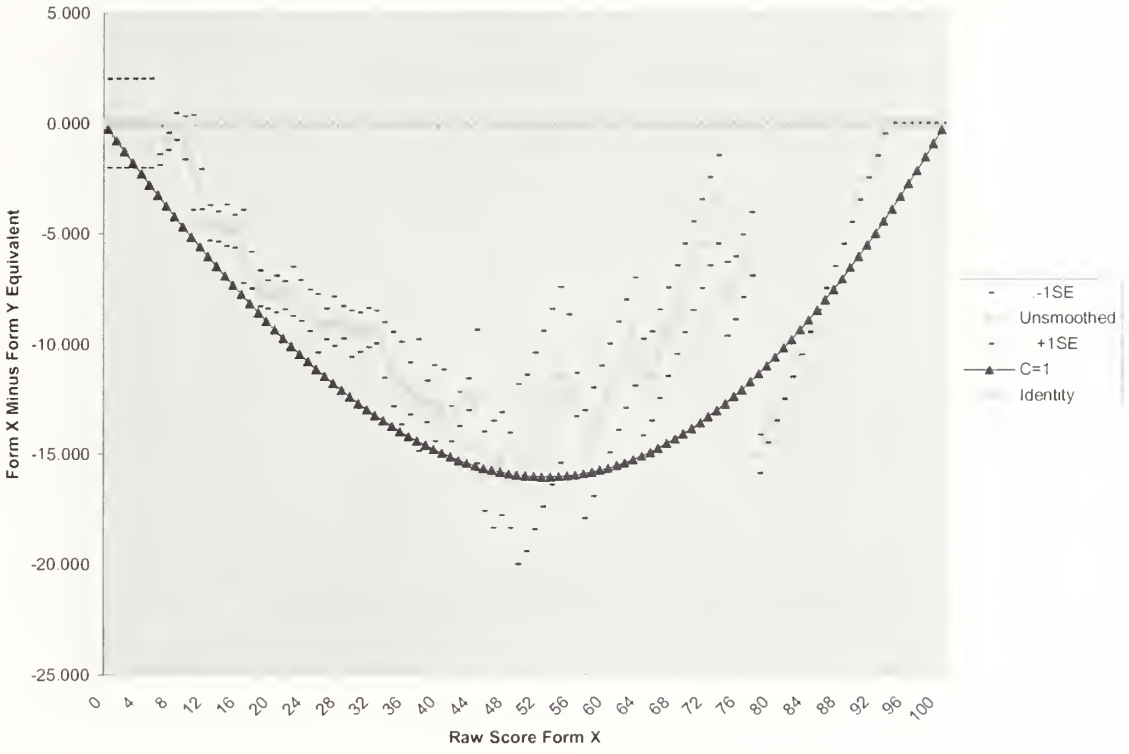

Figure 4.1: Unsmoothed Function versus C = 1 Smoothed Function.

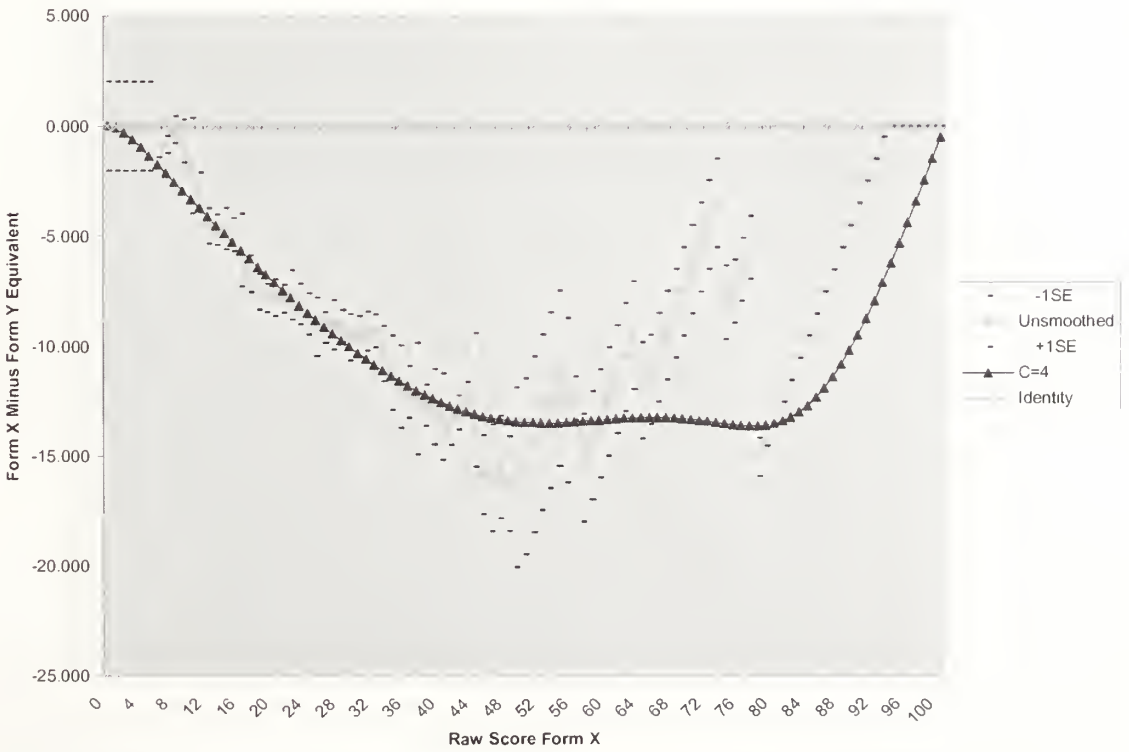

Figure 4.2: Unsmoothed Function versus C = 4 Smoothed Function.

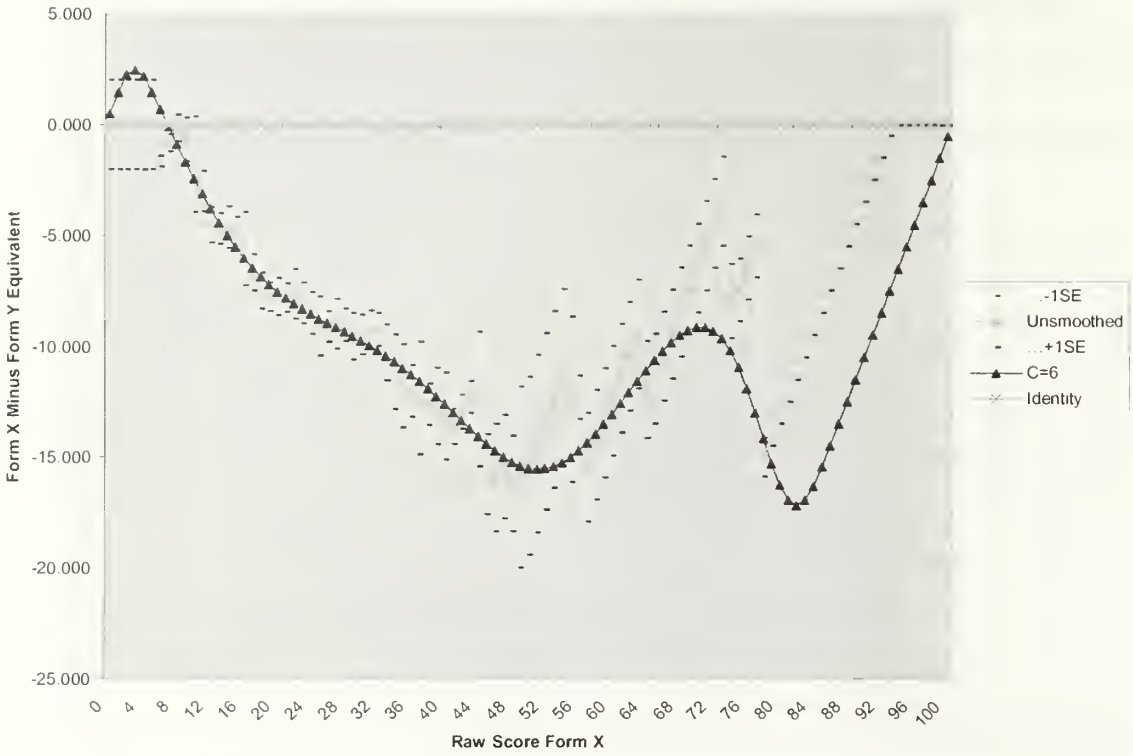

Figure 4.3: Unsmoothed Function versus C = 6 Smoothed Function.

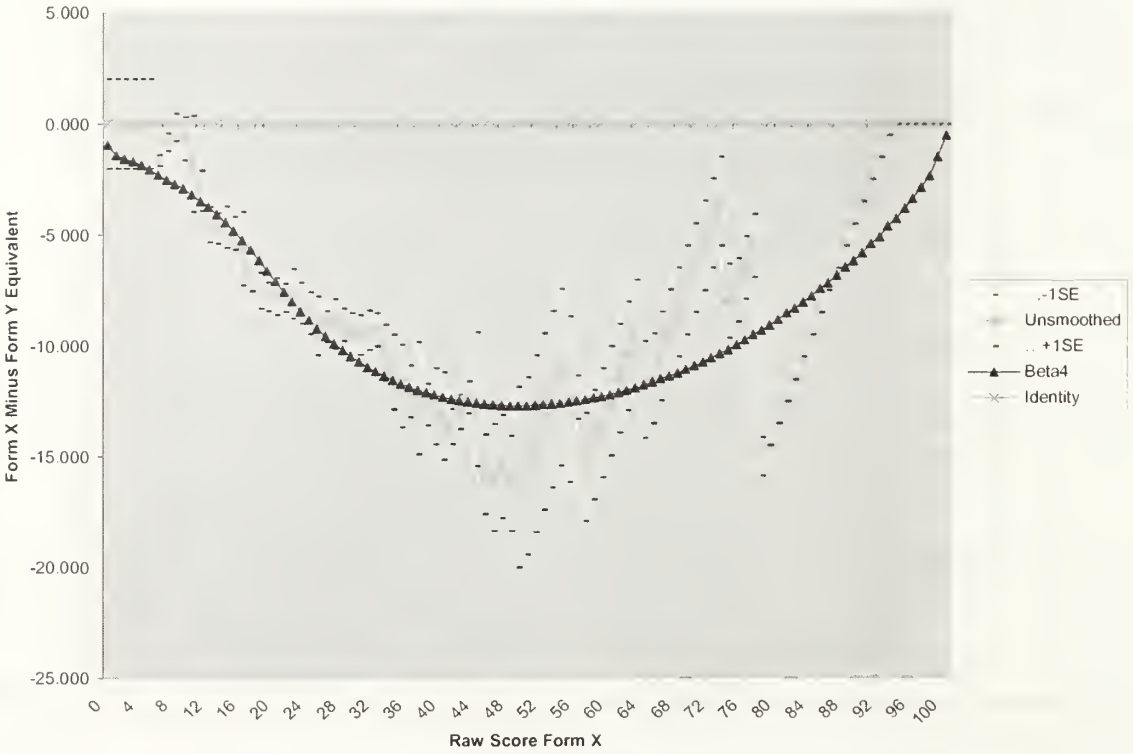

Figure 4.4: Unsmoothed Function versus Beta4 Equating Function.

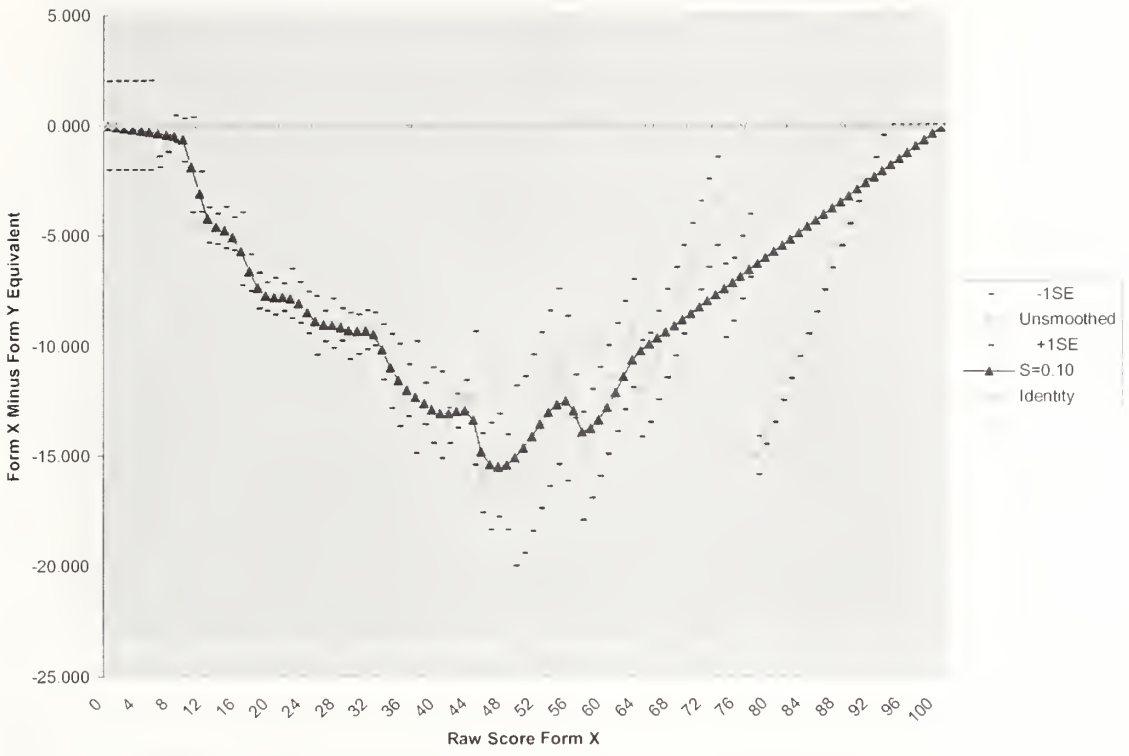

Figure 4.5: Unsmoothed Function versus S = 0.10 Smoothed Function

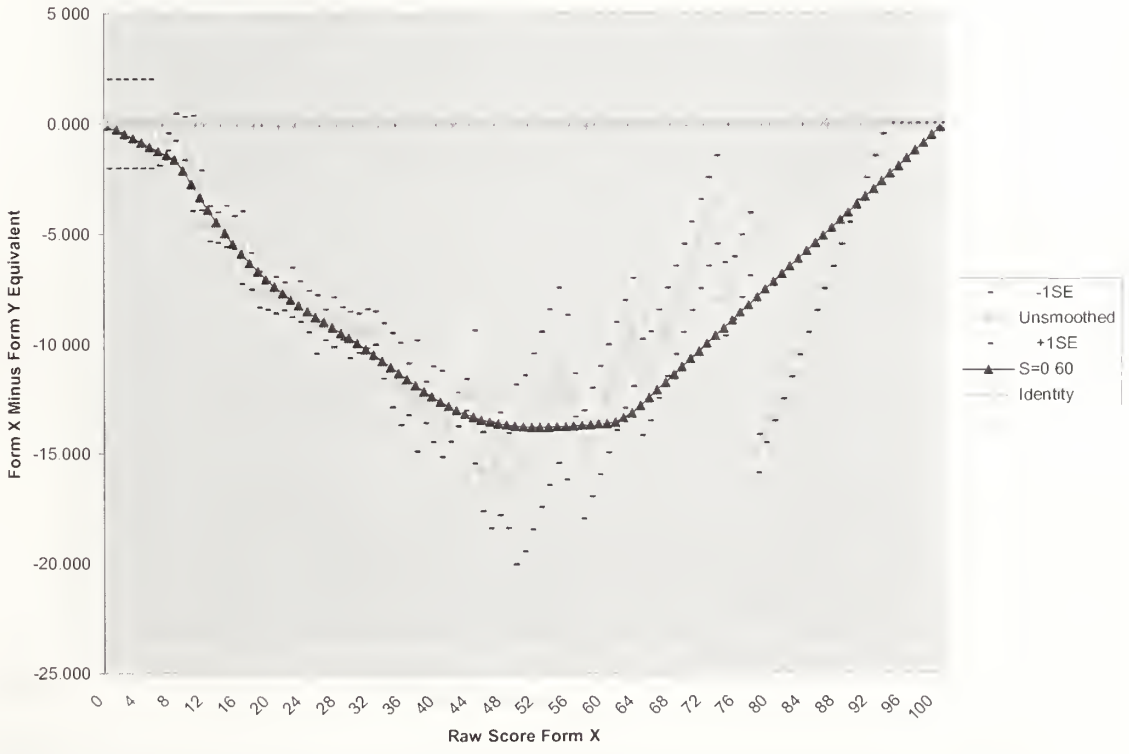

Figure 4.6: Unsmoothed Function versus S = 0.60 Smoothed Function

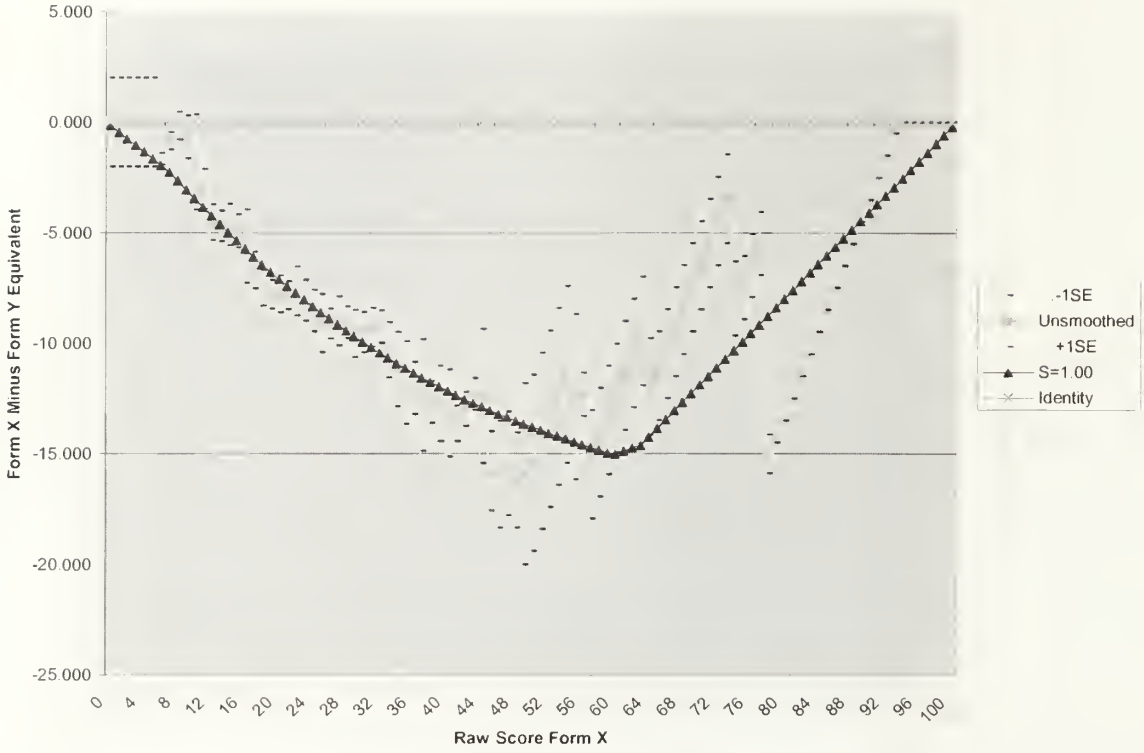

Figure 4.7: Unsmoothed Function versus S = 1.00 Smoothed Function

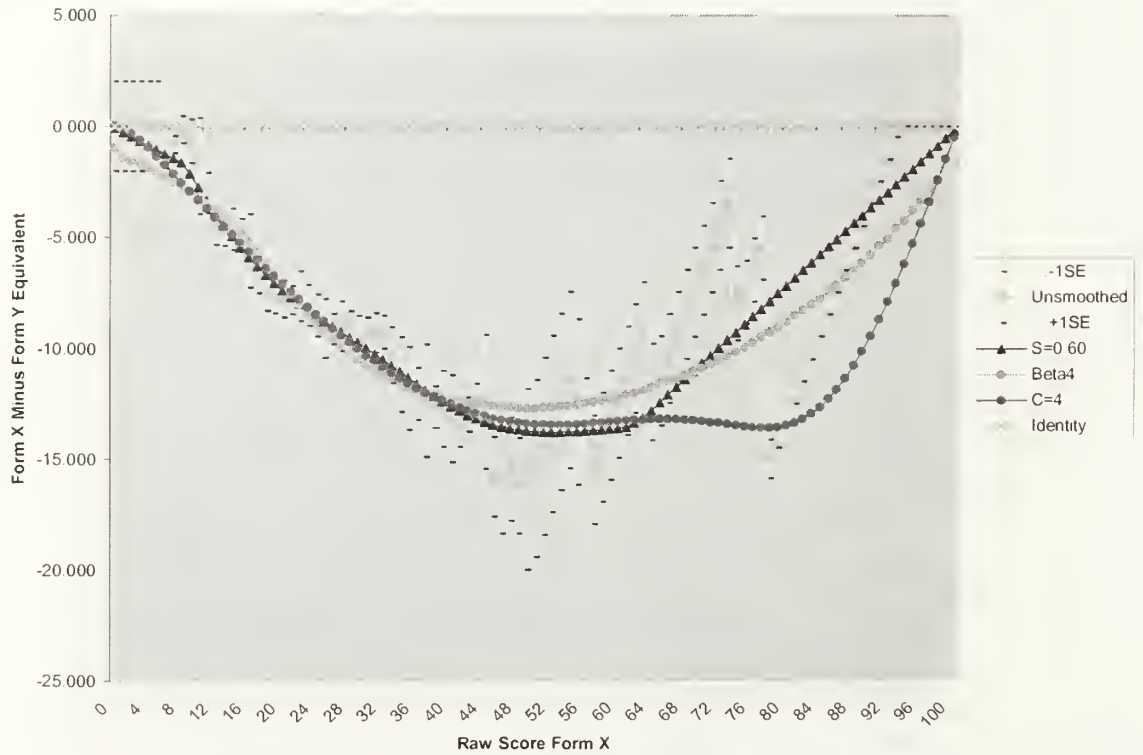

Figure 4.8: Smoothed Distributions for Log-linear (C = 4), Beta4, and Cubic Spline (S = 0.60)

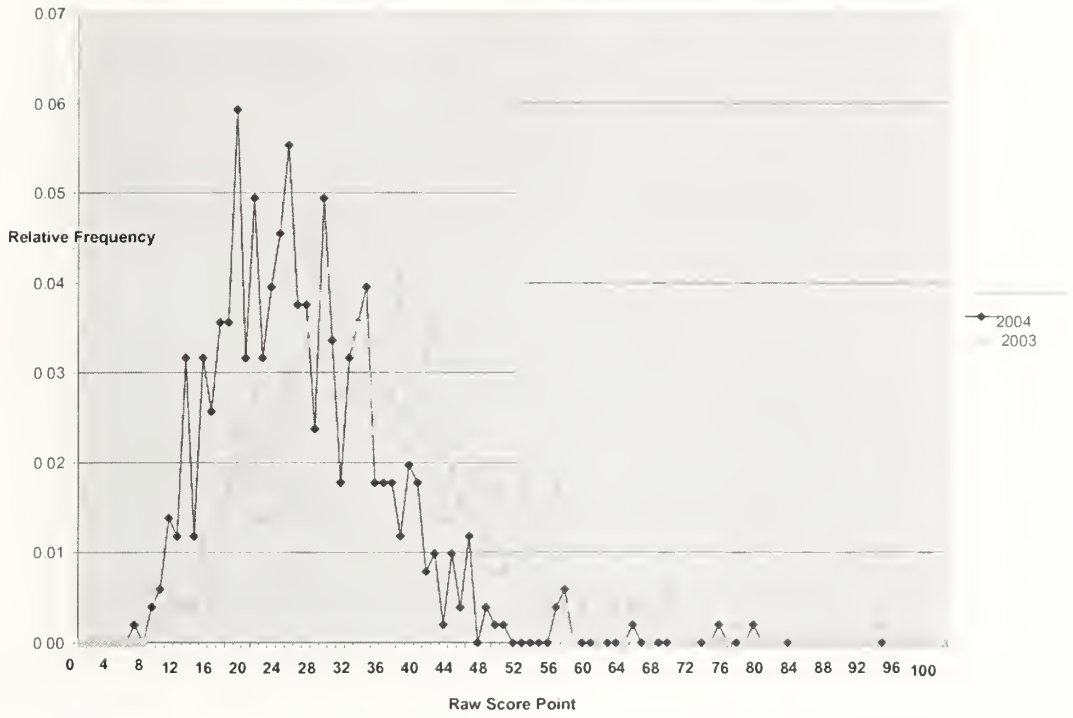

Figure 4.9: Relative Frequency Distributions for 2004 and 2003 Tests

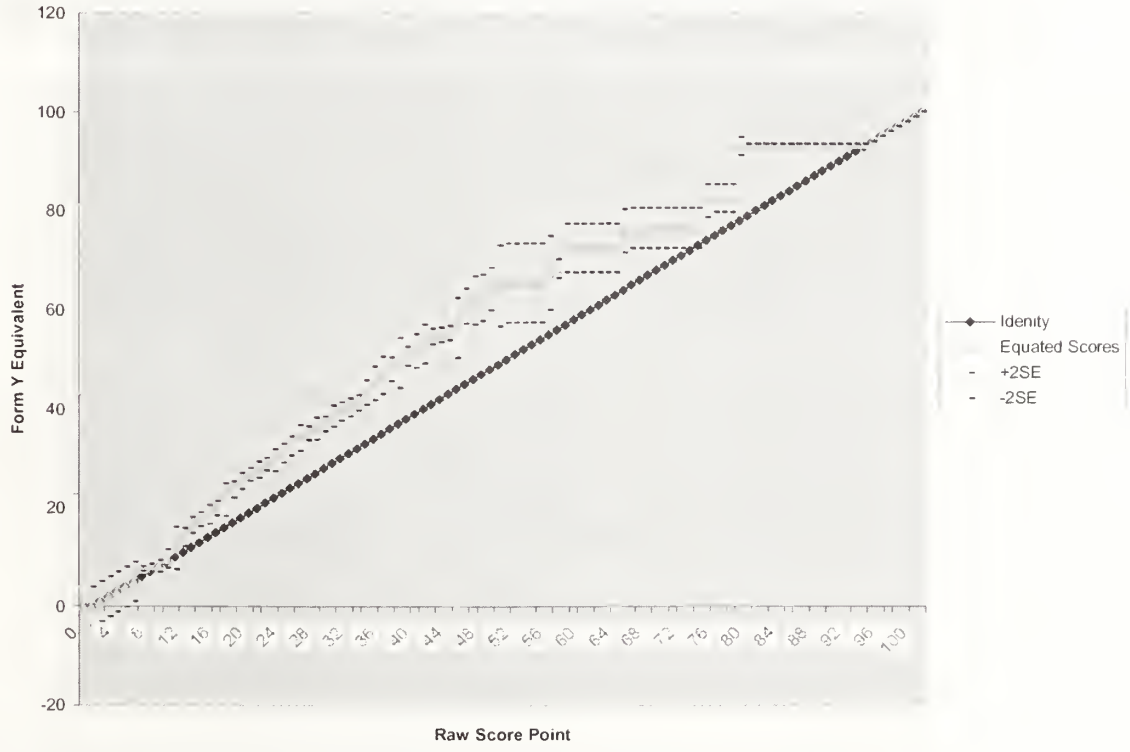

Figure 4.10: Identity versus Equipercentile Equating Relationships

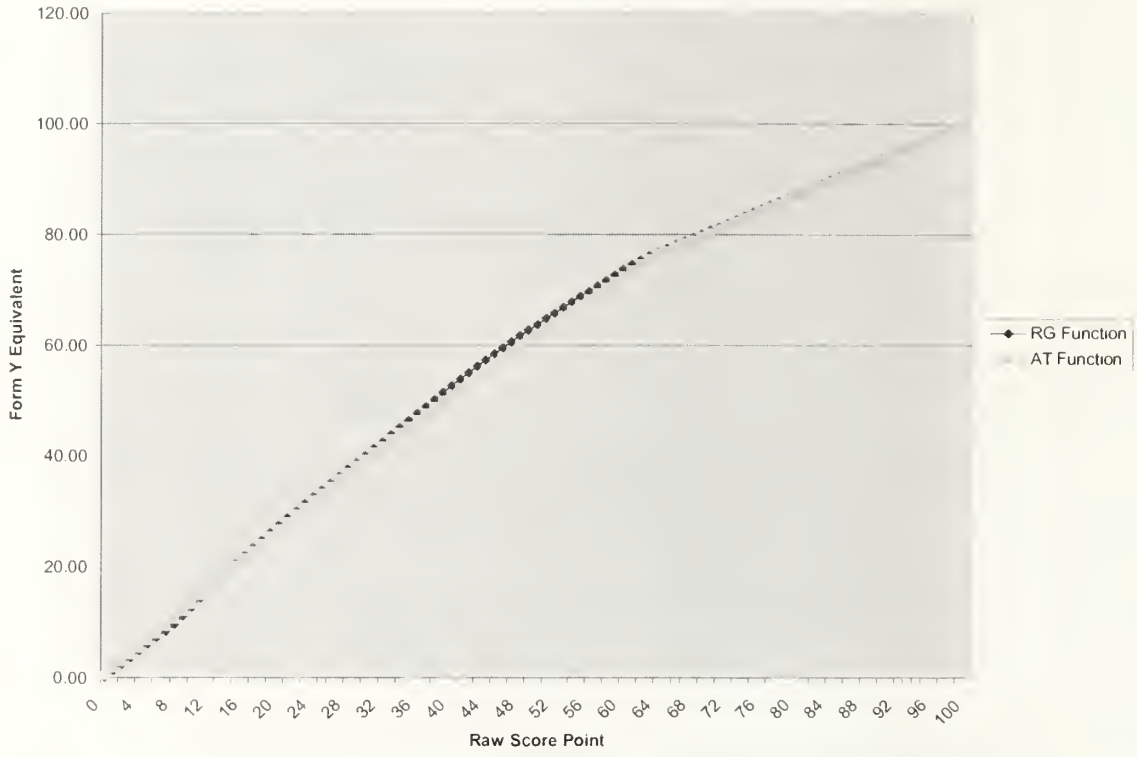

Figure 4.11: Smoothed Random Groups and External Anchor Test Equipercentile Functions of 2004 to 2003 Forms.

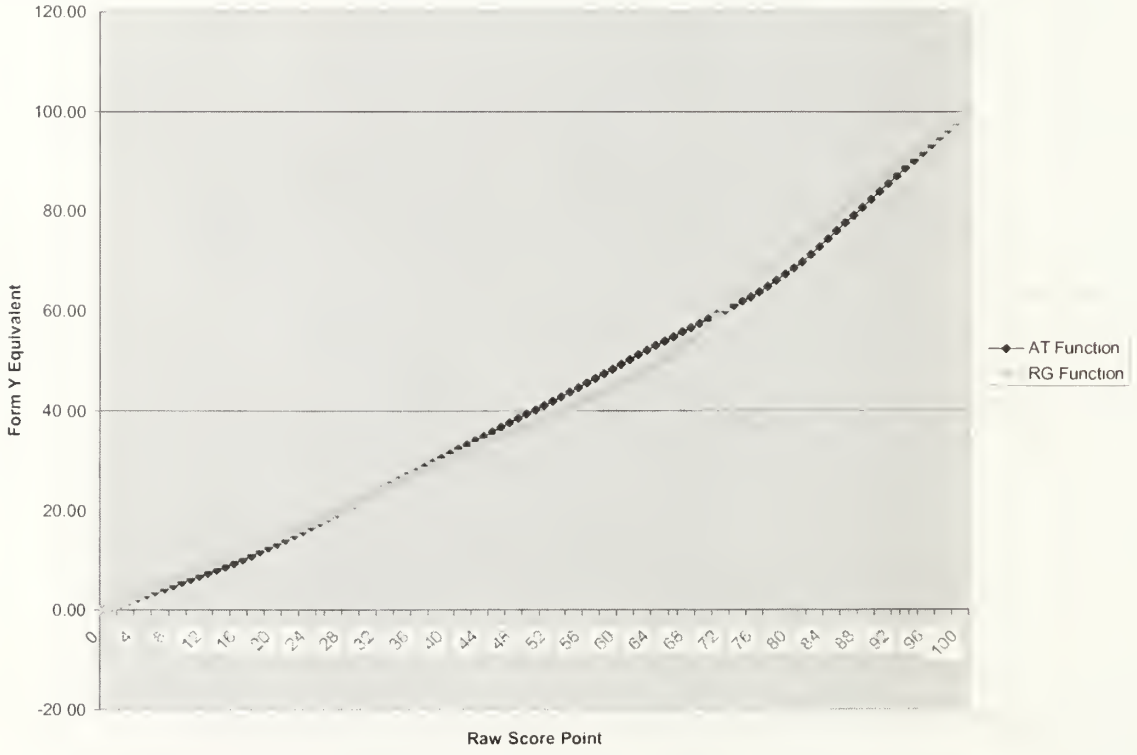

Figure 4.12: Smoothed Random Groups and External Anchor Test Equipercentile Functions of 2005 to 2004 Forms.

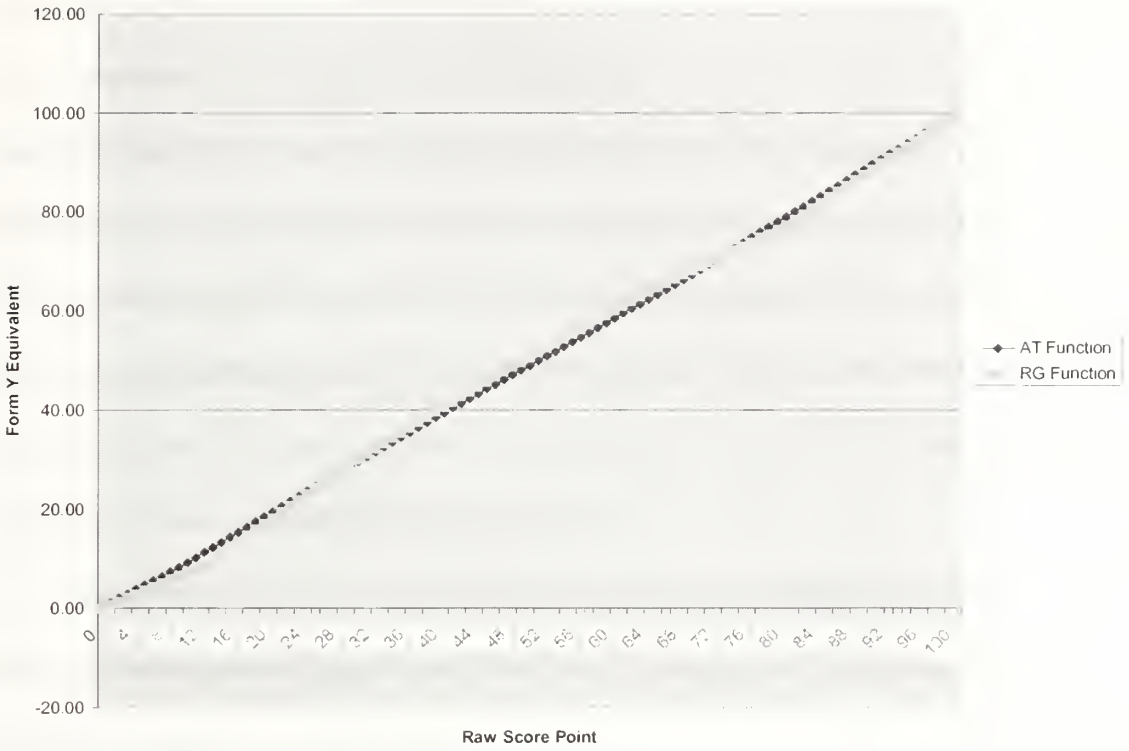

Figure 4.13: Smoothed Random Groups and External Anchor Test Equipercentile Functions of 2005 to 2003 Forms.

CHAPTER 5

SUMMARY OF FINDINGS

In chapter 4 detailed results of the study were reported. This chapter, however, presents a summary of the findings, the significance of the findings, the delimitations of the study and directions for future research. The last section offers some recommendations to MANEB in particular and to other examinations boards with similar practices.

5.1 Summary of Findings

5.1.1 Is it Necessary to Equate MANEB Tests?

Equating has become a household name for most testing agencies in the United States of America, Canada, and some countries in Europe. The practice, however, is not appreciated by some agencies in the United Kingdom and Africa. For MANEB, the practice of not equating educational tests was inherited from earlier British examination systems. As long as the test forms are constructed from the same curriculum benchmarks and are based on the same specifications, they are regarded as similar enough to have their scores compared regardless of the year or occasion they are administered. However, the findings of this study run counter to this line of thought. They show that even if the tests are modeled on the same curriculum and despite the best effort by test constructors to come up with parallel forms, the tests are dissimilar in difficulty and distribution of scores.

The null hypothesis that distributions of raw scores on the 2003 and 2004 tests were similar was rejected signifying that the distributions were not similar enough to warrant equating unnecessary. According to Harris and Crouse, (1993) equating becomes

unnecessary when scores from the test forms to be equated are very similar. The relative score distribution for the 2004 test was more positively skewed, and more leptokurtic than the 2003 score distribution, although both were positively skewed distributions. The study has also empirically shown that equivalent groups of examinees were performing differently on the test forms indicating that the test forms were dissimilar in difficulty. The message from these findings, therefore, is that it is necessary to equate the PSLCE Mathematics test forms developed by MANEB so that their levels of difficulty and their distributions can be matched.

It is appreciated that examinations boards have the expertise to develop technically sound and highly comparable test forms through a very rigorous process. However, these measuring instruments in education and the rest of the social science world usually remain different instruments. Since these are high stakes examinations, fairness demands that a relationship between scores from different test forms should first be defined before making any kind of comparison. Such a relationship can be defined through the process of equating. This process would stretch and compress the scale of one form so that its distribution would coincide with distribution of the other form (Angoff, 1982). As a consequence of this operation, randomly equivalent groups would earn similar converted scores regardless of the form taken.

Results from the identity equating methods employed in this study support this conclusion. When the equated scores were compared to the identity scores, the identity line fell outside the ± 2 standard error band. This finding signified that equated scores were different from the scores that were not equated and that equating was better than not equating at all. According to Kolen and Brennan (2004) such a result indicates that

equating is necessary. Therefore, the PSLCE Mathematics tests developed by MANEB should be equated. It is not enough to use the same curriculum benchmarks and test specifications when developing test forms. Additional processes such as equating are required to make scores across forms comparable and equivalent.

5.1.2 Consequences of Not Equating Educational Tests

Another purpose of this study was to investigate the consequences of not equating educational tests. The consequences of interest were the classification of students into different grade boundaries and the invariance of examination standards across test forms. The idea was to garner empirical evidence regarding the comparability of classification decisions based on scores that are not equated and classification decisions based on equated scores. A similar and perhaps related idea was to comparability examination standards before and after equating.

MANEB transforms raw scores on the test forms to a letter grade scale and specific score boundaries are determined by the "Awards Committee" comprising highly qualified experts. This process is similar to the one followed by the Welsh Joint Education Committee (WJEC) in the United Kingdom. During this scaling process, the difficulty level of the test form relative to the previous forms is taken into consideration when setting cut scores for the different grade boundaries. One would expect, therefore, that this process completely solves the problem of differential test difficulty and that examinees would be fairly evaluated with reference to the same standards as their colleagues who took previous test forms. However, the results of this study suggest otherwise. It appears that the process does not properly accomplish this task and more needs to be done to fairly and accurately classify students taking different test forms.

The study compared the examination standards represented by cut scores across forms. The passing score on 2004 form differed from the passing score on 2003 form by 0.25 standard deviation units. The cut score for a distinction category was 70 on both forms. However, this score was 4.56 standard deviations above the mean of the group on 2004 test, but only 2.94 standard deviations above the mean of the group on 2003 test representing a difference of 1.62 in standard deviation units across the randomly equivalent groups. The cut scores for other categories (B, C, and D) were all different in standard deviation units from their respective group means. Therefore, the standards used to judge the performance of examinees were different across forms in spite of the best effort during the Awards Meeting to set equivalent standards. These differences affected the way students were classified into grade boundaries. Classifying students using such standards could be both unfair to examinees and it could provide misleading information to policy makers.

For example, the pass rates on 2004 (69.96%) and 2003 (81.41%) test forms were different. Just by looking at these pass rates, school administrators, policy makers, and the public would conclude that students did better on 2003 test than on 2004 test, which is correct. They would presumably go further to allege that students who took 2003 test were brighter than those who took 2004 exam, which is incorrect. In fact, the two groups were randomly equivalent with respect to the construct the tests were measuring. The difference in pass rates would also misinform stakeholders that the standards of education were plummeting. Perhaps, costly reforms would then be conceived and initiated in the school system to try to redeem the situation. The correct explanation, however, is that 2004 test was more difficult than 2003 test and although this information was considered

during the "Awards Meeting," the cut scores were not good enough to eliminate the differences. The observed differences in the classification of students into different grade boundaries shown in this study were neither as a result of one group being brighter than the other group, nor was it because of the plummeting educational standards. Rather the differences arose, among others, from the differences in test difficulty and varying examination standards. Therefore, one important message from this finding is that the process that is used to set cut scores needs to be improved or changed because it does not guarantee invariance of examination standards.

In the interest of fairness to students taking different test forms and also to better inform stakeholders about the performance of different cohorts of students, it is important to rectify the problems noted in this section. It is important to adjust for difficulty scores on tests forms and to employ a better way of maintaining the same examination standards across forms. One procedure that would help solve these problems is to begin to equate test scores. After equating, the difference in pass rates on 2004 test and on 2003 test reduced from 11.54% to 7.33%. The pass rate on 2004 test was now 88.73% whereas that on 2003 reference test form remained 81.41%. By equating scores in this instance eradicated the problem of differences in test difficulty and helped to bring the 2004 pass rate closer to the 2003 pass rate than before. The proportion of false negative examinees on 2004 test was reduced. However, the problem of differences in cut scores still remained.

This problem could also be solved by using the same cut scores on both forms and there are two ways these cut scores could be determined. One way would be to have stakeholders agree that whatever the distribution of scores, the passing score on any test

form would be set at a raw score point that would allow for a certain percentage of examinees (e.g., 80% of the examinees) to pass the exam. Such cut scores expressed in terms of the number or percent of examinees passing are referred to as "relative standards" and they are appropriate for selection and placement examinations. If this procedure is adopted, the pass rate on every test form would be the same. The downside to this procedure, however, is that it would be almost impossible to measure growth and there would be no way of knowing whether or not the educational system is changing. More importantly, the cut scores determined in this way do not necessarily have meaning. Alternatively, MANEB can set "Absolute Standards" usually determined through standard setting procedures.

Absolute standards are expressed as number or percentage of correct responses on a test form (e.g., 40 correct responses of the 100 items – 40%). The process usually begins with development of general descriptions for each performance category. Then content specialists can develop performance level definitions for each subject area. After these descriptions are approved by authorities, standard setting procedures such as the Angoff's, Contrasting Groups, Body of Work, Bookmark, and Item-Mapping methods may be employed to determine specific cut scores. Using absolute standards determined in this way enables stakeholders to measure grow and monitor changes in the education system. Furthermore, such cut scores have meaning because they are based on performance descriptors defining what examinees must know or be able to perform for them to be regarded as passing. Because of the rigorous nature of the standard setting procedures, it is usually less cost-effective to set cut scores on every test form that is

developed. Consequently, standards are usually set on the reference form of the test only and they are usually maintained across forms through equating.

In this study, when the passing score on the reference 2003 test form (raw scores point of 24) was maintained across forms, the pass rate on the 2004 test changed to 82.61% whereas the pass rate on the 2003 form was 81.41% reducing the difference further to 1.2% after equating. Note that maintaining the same cut scores across test forms can only be done when scores are equated because before equating scores on the two forms are not equivalent and it would be unreasonable to pick a cut score for one form and use it on the other form. With these results, it is now easy to tell that the pass rates were not different on the two test forms. In fact, they were now consistent with what was expected since the two groups that took the tests were equivalent in terms of mathematics ability. Stakeholders would now be well served with these results. What looked like the problem of dwindling standards of educations has turned out to be the problem of the way test scores were treated before reporting. Similar findings were observed when the classifications of students into different grade boundaries were compared before and after equating. The differences were smaller after equating and they became even much smaller when the same cut scores were maintained across forms. These results signify that equating can be very instrumental in promoting fairness and in facilitating accurate score reporting to stakeholders.

Maintaining cut scores through equating as demonstrated in this study would serve to meet the expectation of the public and the often desirable goal for many assessment systems. The public believes that the passing mark on exams does not change and in the interest of fairness, they do not expect it to change. With equating, this

expectation can easily be satisfied and professional conspiracy theories would easily be contained. The process would also render meaningless the policy of not publicizing cut scores, which in turn would help to boost the integrity of the examination system. Furthermore, maintaining the same cut scores across forms through equating would eradicate the need to hold "Awards Meetings" every year, a cost effective way of saving both time and resources.

5.1.3 Equating Using an External Anchor Test

The third purpose of the study was to explore whether educational tests can be adequately equated using an external anchor test. The investigation centered on learning the usefulness of the external anchor test that is administered separately from the operational test. In this study the anchor test was administered 5 weeks before the operational 2005 PSLCE mathematics test. The 2004 to 2003 equating relationship obtained using the external anchor test design was compared to the 2004 to 2003 equating relationship obtained using the random groups design. In this study, the external anchor test design worked just as well as the random groups design.

The first part of the investigation was to compute the reduction in uncertainty indices for the test forms and in all cases they were small. This finding signified that the external anchor test could not provide enough information to reduce the uncertainty about the test forms. However, the RIU index depends on the degree of linear relationship between the anchor test and the test forms and the small magnitude of the reported indices was as a result of small magnitude of the correlations between the anchor test and the test forms. It is conceivable that the tests could be related in a curvilinear way or in some other ways which were not investigated in this study. Furthermore the small

correlations of the anchor test and the test forms observed in this study were consistent with the observations in earlier studies (Kolen, 1991; Liou, Cheng, & Li, 2001). Since the length of the anchor test is usually short, the correlation between the anchor and the test form is oftentimes low. Nevertheless, developing an external anchor test that is highly correlated with test scores would help to increase the magnitude of the RIU, which in turn would render the usefulness of the anchor test more defensible.

The other findings were consistent with the conclusion that the external anchor test worked well. The mean squared equating errors for the random groups design were comparable in magnitude to the mean squared equating errors for the external anchor test design except for the Tucker equating of the 2004 test to the 2003 form. More importantly, the standardized root mean square difference between the random groups and the external anchor test equating functions was small in all cases. In fact, plotting the functions on the same axes as displayed in Figures 4.11, 4.12, and 4.13 showed that the lines get superimposed on each other except in a few points on the score scale. These results support the conclusion that the external anchor test design was as useful as the random groups design in equating of the test forms. The small magnitude of the mean equating errors also signifies that equating was adequate.

When scores on the test forms were equated through the anchor test using both equipercentile and linear equating methods and the classification of students into pass/fail categories were compared before and after equating, the results showed that equating was better than not equating at all. The classifications were more comparable after equating than before equating except for the Tucker linear equating of the 2005 test to the 2003 test form where the classification after equating did not change. The proportion of false

positive examinees (i.e., examinees that were classified as passing when, in fact, they should have been classified as failing) was highly reduced after equating in all cases. Therefore, even if the external anchor test were less useful by RIU determination, equating through the anchor would still help to make better classification decisions than identity equating.

One major concern in the external anchor test design proposed in this study would be the presence of learning effects from the time students take the anchor to the time they take the target test. The fact that the difference between the random groups and external anchor test designs was small and yet the external anchor test was administered 5 weeks before the operational test provided evidence that the learning effects between the administrations were minimal. In fact, comparison of the group means on 2004 and 2005 and also on 2003 and 2005 test forms through the paired sample t-tests, suggested that learning effects could be ruled out. This finding was very important in that it supported the idea that the external anchor test can be administered 5 weeks before the target tests and still not be confounded by learning effects.

The most important lesson arising from these findings is that the external anchor test design presented in this study ought to be considered as a modification of the popular and perhaps more realistic design, the non-equivalent groups with anchor test (NEAT) design. The NEAT design is used by most testing agencies in USA, but it could not directly be used by examinations boards like MANEB without modifications. The design requires the use of anchor items that are not exposed, which would be a problem with MANEB where all the items are released after each administration. There are also concerns of over burdening examinees by administering the anchor test together with the

operational test especially for high stakes examinations. Therefore, administering an external anchor test separately from the operational test as shown in this study would provide an alternative modification of the NEAT design. Furthermore, examinations board like MANEB sometimes field test the items before assembling them into an operational test and as such the board can take advantage of this exercise to administer the anchor test to a random sample of examinees provided the time between the administrations is kept short to minimize the learning effects. Certainly such a plan would be cost effective since it combines two exercises in one operation. To ensure that students remain motivated, it would be a good idea to administer the anchor test (or conduct pilot-testing) when schools are already carrying out their mock (practice) exercises as demonstrated in this study. The scores on the anchor given to a random sample of students during field testing would be used to generate a conversion table, which would then be used for the population of students.

The external anchor test design investigated in this study has other advantages too over the traditional NEAT design mentioned in the previous section. In cases where all the items are exposed after administration, teachers can construct the items after some training as was the case in this study. In this instance, it is not necessary that the items comprising the anchor test be part of the reference form and it is not necessary that they be reused in the next administration. In any case, however, it may still be better and cost effective to maintain the same set of anchor test for some time. The fact that the external anchor test is given separately from the target test may make it easy for the Board to collect back the question papers after administration. Such an exercise is difficult to accomplish when the anchor and operational tests are given concurrently.

5.2 Significance of the Findings

The significance of the results presented in this dissertation stems from the argument-based approach the study took to garner evidence relating to the practice of equating. The MANEB, policy makers, the media, educators, researchers, and the public in Malawi continue to compare examination results across test forms even when those results were not equated. The information in this study should guide these stakeholders regarding how far they can interpret the test score that are not equated. In this regard, information will form part of the validity evidence to support the PSLCE mathematics test score use and interpretation.

The major contribution of this study is that it has succeeded in making a case for equating. It is hoped that the findings of the study will inform dialogue within MANEB and in other similar examinations boards to seriously consider equating test forms to complement their already sound test development and score reporting processes. Ignoring equating has negative consequences which affect major decisions they make about examinees, and it affects the quality of information they disseminate to the public and to policy makers about what is going on in schools. This study conveys the important message that the quality of the decisions made is only as good as the quality of information on which they are based. Equating can help to improve the quality of the information MANEB offers the public.

Finally, the study has shown that the external anchor test design that is administered separately from the operational test works well enough and its outcomes were as good as those obtained from the random groups designs. The fact that the random groups design has remained a scientifically robust procedure for collecting data for

decades, the comparability of the external anchor test design to this widely acceptable design is good news for the Board. Therefore, this information should provide a starting point to MANEB and similar examinations boards as they search for proper equating designs for their tests.

5.3 Delimitations and Direction for Future Research

The findings of this study are limited by several factors. MANEB develops and administers numerous tests and this study has only looked at one subject test, the PSLCE mathematics test, and even for this subject, only three test forms were involved. While it can be argued that these tests are similar in many ways including the way they are developed, administered, and have their cut scores set, the findings of this study should be cautiously generalized. It is important to have further investigations using other subject test forms. For example, other studies may investigate the consequences of not equating the Malawi School Certificate of Education (MSCE) tests forms and compare the results to the findings of this study. The new studies may also include more test forms to span a period of more than three years and the replicability of these results across many forms would add to its existing reliability.

The schools participating in this study were all drawn from Zomba district. An attempt was made to draw a representative sample of schools from the district and all types of schools were captured. However, this is just one district among 27 administrative districts in Malawi and it is difficult to tell whether or not its characteristics are representative of the characteristics of all the districts in the country. Although it can be argued that the population of schools in Zomba is typical of the population of schools in a majority of the districts, the geographical factors are different and these differences in

geographical factors render the population of examinees different across districts. Therefore, conducting the same study in other districts or with a representative sample of the districts may yield different results.

The sample size used in the study was small such that standard errors of equating at some score points in all cases were bigger than 0.10. Larger samples for each test form are required for standard error of equating at any score point to be less than 0.10 (Kolen & Brennan, 2004). More studies with larger samples, therefore, are needed to investigate the same problems as in this study. This would help to properly evaluate the adequacy of equating using the external anchor test design and provide further evidence regarding the comparability of the external anchor test design and the random groups design.

Equating requires that exposed items should not be used in the process. Students in this study may have already seen the test items on the 2004 and the 2003 test forms. Although an attempt was made to collect such evidence, teachers and students alike may have provided socially desirable responses. In fact, such responses were observed when teachers were asked to help identify students who were repeating the grade. Therefore, further studies ought to be conducted using test forms that are administered for the first time. Such studies would perhaps yield different results.

The study has shown that the external anchor test design provides comparable results to the random groups design. While these results are useful, the study should be viewed as a foundation upon which the Board and other researchers may build their future work in search for a suitable equating design. Further research should try to replicate the findings of this study before using it in any high stakes testing program. Replicating the study may be a good idea because of the low reliabilities of the test forms

used in this study. There are many reasons that may explain the low reliabilities observed for 2004, 2003, and anchor tests. The tests were difficult for the students participating in the study and as such it is possible that many students were just guessing the right responses. Furthermore, the content areas assessed by the test were so heterogeneous that it would be difficult to reliably assess each one of them. This content heterogeneity, in turn, may have affected the reliability of the test form. Although no major problem was encountered during scoring, the inter-rater reliability was not computed and as such the consistency with which the scripts were scored is not exactly known. Therefore, scoring may be one factor as well that might have contributed to low reliabilities of the test forms. Therefore, more studies using MANEB data with high reliability may provide different results than those reported in this study.

5.4 Recommendations

In light of the findings of this study while taking into consideration its limitations, the following recommendations can be made to MANEB and the recommendations may also be helpful to similar examinations boards that do not equate test score.

1. The current practice by MANEB of not equating educational tests should be reconsidered because, contrary to popular belief, when equating is not implemented, the distribution of scores and the test difficulties are not appropriately adjusted and so a student's likelihood of passing is affected by the test form she or he took.
2. MANEB should consider maintaining the same cut scores across test forms through equating because the current process used in setting the standards fails, in spite of the best effort, to set comparable cut off points. This weakness makes the

invariance of the standards of examinations indefensible and it leads to candidates taking different versions of the test to be evaluated with reference to different standards.

3. MANEB, policy makers, the media, educators, researchers, and the public should desist from making direct comparison of results on test forms administered across years, unless the scores are equated. The comparison of such scores, even in terms of the pass rates or classification of students into grade boundaries, tells us nothing about whether one cohort is better or worse than the other.
4. MANEB should also consider equating its operational test forms using the external anchor test design as a modification of the NEAT design provided the time interval between the administrations should be short. This design can help to curtail the fear of over burdening examinees during the administration of the operational tests and also it can provide a leeway where all items are exposed after each administration.
5. MANEB may administer the anchor test to a random sample of students, instead of the whole population of students in a particular year. The data collected from the random sample would then be used in the equating process and such an exercise would be economical to MANEB.
6. For purposes of capturing students' motivating, the exercise may take place as part of the practice (mock) examinations as was the case in this study. This too would help to minimize the time interval between the administration of the anchor test and the administration of the target test form.

REFERENCES

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Angoff, W.H. (1971). Scales, norms, and equivalent scores. In R.L. Thorndike (Ed.), *Educational Measurement* (2nd ed., pp. 508 – 600). Washington, DC: American Council on Education. (Reprinted as W.H. Angoff, Scales, Norms, and Equivalent Scores. Princeton, NJ: Educational Testing Service, 1984).

Angoff, W.H. (1987). Technical and practical issues in equating: A discussion of four papers. *Applied Psychological Measurement*, 11, 291 – 300.

Angoff, W.H., & Cowell, W.R (1986). An examination of the assumption that the equating of parallel forms is population independent. *Journal of Educational Measurment*, 23, 327 – 345.

Assessment and Qualification Alliance. (2004). *Uniform marks in GCE, VCE, GNVQ and GCSE examinations*. Retrieved February 17, 2006, from http://www.aqa.org.uk/over/stat.pdf/uniformmarks-leaflet.pdf

Braun, H.I., & Holland, P.W. (1982). Observed score equating: A mathematical analysis of some ETS equating procedure. In P.W Holland and D.B Rubin (Eds.), *Test equating* (pp. 9-49). New York: Academic.

Budescu, D. (1987). Selecting an equating method: Linear or equipercentile? *Journal of Educational Measurement*, 22, 13-20.

Crocker, L., Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart, and Winston.

Cui, Z. & Kolen, M.J. (2005). *RAGE-RGEQUATE [Computer Program]*. Iowa City, IA: The University of Iowa, Iowa Testing Programs.

Dorans, N.J. (1990). Equating methods and sampling designs. *Applied Measurement in Education*, 3, 3 – 17.

Dorans, N.J. (2002). Recentering and realigning the SAT score distributions: how and why. *Journal of Educational Measurment*, 39, 59 – 84.

Dorans, N.J. (2004). Equating, concordance, and expectation. *Applied Psychological Measurement*, 28(4), 227 – 246.

Dorans, N.J., & Holland, P.W. (2000). Population invariance and equitability of tests: Basic theory and the linear case. *Journal of Educational Measurment*, 37, 281 – 306.

Dorans, N.J., & Lawrence, I.M. (1990). Checking the statistical equivalence of nearly identical test editions. *Applied Measurement in Education,* 3, 245- 254.

Eignor, D.R., Stocking, M.L., & Cook, L.L. (1990). Simulation results of effects on linear and curvilinear observed- and true-score equating procedures of matching on a fallible criterion. *Applied Measurement in Education*, 3, 37 – 52.

Fairbank, B.A., Jr. (1987). The use of presmoothing and postsmoothing to increase the precision in equipercentile equating. *Applied Psychological Measurement,* 11, 245 – 262.

Flanagan, J.C. (1951). Units, scores, and norms. In E.F. Lindquist (Ed.), *Educational Measurement* (pp. 695 – 763). Washington DC: American Council on Education.

Gafni, N., Melamed, E. (1990). Using circular equating paradigm for comparison of linear equating methods. *Applied Psychological Measurement*, 14, 243 – 256.

Gulliksen, H. (1950). *Theory of mental tests*. New York: Wiley.

Hanson, B., & Cui, Z. (2004). *Equating error [computer program*]. Iowa City, IA: The University of Iowa, Iowa Testing Programs.

Harris, D.J., & Crouse, J.D. (1993). A study of criteria used in equating. *Applied Measurement in Education,* 6, 195 - 240.

Harris, D.J., & Kolen, M.J. (1990). A comparison of two equipercentile equating methods for common item equating. *Educational and Psychological Measurement*, 50, 61-71.

Holland, P.W. & Rubin, D.T. (1982). *Test equating*. New York: Academic Press.

Jarjoura, D., & Kolen, M.J. (1985). Standard errors of equipercentile equating for the common item nonequivalent populations design. *Journal of Educational Measurment,* 10, 143 – 160.

Keeves, J.P. & Alagumalai, S. (1999). New approaches to measurement. In G.N. Masters & J.P. Keeves (Ed.), *Advances in Measurement in Educational Research And Assessment,* (pp. 39 – 40). An imprint of Elsevier Science, Amsterdam: Pergamom.

Kelley, T.L. (1923). *Statistical methods*. New York: Macmillan.

Klein, L.W., & Jarjoura, D. (1985). The importance of content representation for common item equating with nonrandom groups. *Journal of Educational Measurement,*22, 197 – 206.

Kolen, M.J. (1985). Standard errors of Tucker equating. *Applied Psychological Measurement*, 9, 209 – 223.

Kolen, M.J. (1988). Instructional topics in educational measurement. *Educational Measurement: Issues and Practice*, Winter, 29 – 36.

Kolen, M.J. (1991). Smoothing methods for estimating test score distributions. *Journal of Educational Measurement*, 28(3), 257 – 282.

Kolen, M.J. (1999). Equating of tests. In G.N. Masters & J.P. Keeves (Ed.), *Advances in measurement in educational research and assessment* (pp. 164 - 175). An Imprint of Elsevier Science, Amsterdam: Pergamom.

Kolen, M.J. (2004). Linking assessment: Concept and History. *Applied Psychological Measurement*, 28(4), 219 – 226.

Kolen, M.J. (2003). *Common item program for equating (CIPE) [Computer Program*]. Iowa City, IA: The University of Iowa, Iowa Testing Programs.

Kolen, M.J. & Brennan, R.J. (1995). *Test equating: Methods and practices*. New York: Springer-Verlag.

Kolen, M.J. & Brennan, R.J. (2004). *Test equating, scaling, and linking: Methods and practices*. New York: Springer-Verlag.

Kolen M.J., & Harris, D.J. (1990). A comparison of item preequating and random groups equating using IRT and equipercentile methods. *Journal of Educational Measurement*, 27, 27 – 39.

Linn, R.L. (1993). Linking results of district assessment. *Applied Measurement in Education*, 6, 83 – 102.

Liou, M., Cheng, P.E., & Li, M. (2001). *Estimating comparable scores using surrogate variables*. Applied Psychological Measurement, 25(2), 197 – 207.

Little, R.J.A., & Rubin, D.B. (1987). *Statistical analysis with missing data*. New York: Wiley.

Livingston, S.A. (2004). *Equating test scores*. Princeton, NJ: Educational Testing Service.

Lord, F.M. (1955a). Equating test scores: a maximum likelihood solution. *Psychometrika*, 20, 193 – 200.

Lord, F.M. (1980). *Application of item response theory to practical testing problems.* Hillsdale, NJ: Erlbaum.

Lord, F.M. (1982). The standard error of equipercentile equating. *Journal of Educational Statistics*, 7, 165 – 174.

Lord, F.M. & Novick, M.R. (1968). *Statistical theories of mental testing scores.* Reading, MA: Addision Wesley.

Morris, C.N. (1982). On the foundations of testing equating. In P.W. Holland & D.R. Rubin (Eds.), *Test Equating* (pp. 169 – 191). New York: Academic Press.

Ogasawara, H. (2001). Standard errors in item response theory equating/linking by response function methods, *Applied Psychological Measurement*, 25(1), 53-67.

Otis, A.S. (1922). The method for finding the correspondence between scores in two tests. *Journal of Educational Psychology*, 13, 529 – 545.

Petersen, N.S., Cook, L.L., & Stocking, M.L. (1983). IRT versus conventional equating methods: A comparative study of scale stability. *Journal of Educational Measurement*, 8, 137 – 156.

Skaggs, G. (1990). To match or not to match samples on ability for equating. A Discussion of five articles. *Applied Measurement in Education*, 3, 105-113.

Thorndike, E.L. (1922). On finding equivalent scores in tests of intelligence. *Journal of Applied Psychology*, 6, 29 – 33.

van Davier, A.A., Holland, P.W. & Thayer, D.T. (2004). *Statistics for social science and public policy: The kernel method of test equating.* New York: Springer-Verlag.

Wang, T, Hanson, B.A., & Harris, D.J. (2000). The effectiveness of circular equating as a criterion for evaluating equating, *Applied Psychological Measurement*, 24(3), 195 – 210.

Wingersky, M.S., Cook, L.L., & Eignor, D.R. (1987). *Specifying the characteristics of linking items used for item response theory item calibration* (Research Report 87 – 24). Princeton, NJ: Educational Testing Service.

Wright, N.K., & Dorans, N.J. (1993). *Using the selection variable for matching or equating* (Research Report No. 92 – 3). Princeton NJ: Educational Testing Service.

Yang, W-L. (2004). Sensitivity of linking between AP multiple-choice scores and composite scores to geographical region: An illustration of checking for population invariance. *Journal of Educational Measurment*, 41(1), 33 – 41.

Zwick, R. (1991). Effects of item order and context on estimation of NEAP Reading Proficiency. *Educational Measurement: Issues and Practice*, 10, 10 – 16.